Origins and Control of Single-Cell Transcript Heterogeneity

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Nicolas Battich

aus

Italien

Promotionskomitee

Prof. Dr. Lucas Pelkmans (Vorsitz)

Prof. Dr. Damian Brunner

Prof. Dr. Josef Jiricny

Prof. Dr. Felix Naef

Zürich, 2016

Origins and Control of Single-Cell Transcript Heterogeneity

Nico Battich, Zurich 2016

Table of Contents

1. Abstract

A central question in biology is the extent to which stochastic molecular processes confine or affect deterministic regulation and the variability between genetically identical cells. Hence, elucidating mechanisms that allow single cells to robustly adjust their phenotypes according to their microenvironment is key to understanding single-cell variability. The composition of the plasma membrane, for instance, is adapted to local cell density through focal adhesion kinase sensing and downstream transcription regulation. The variability of transcript abundance and localization in the cytoplasm of single human cells depicts the main focus of my thesis. I developed image-based transcriptomics, an RNA Fluorescence *in situ* hybridization (FISH) method using branched DNA technology that allows highly reproducible quantification of transcripts at single-molecule resolution. Using image-based transcriptomics, combined with novel image analysis methods and extraction of multivariate feature sets, I determined the subcellular patterning of transcripts and quantified their cell-to-cell variability. Further, I demonstrate that the localization pattern of a transcript correlates with the function of the gene. Although variability of cytoplasmic transcript abundance is large, it is for most genes minimally stochastic, and can be predicted with multivariate models of the phenotypic state and population context of single cells. With computational modeling and experimental validation, I revealed that nuclear retention of transcripts and their export from the nucleus is central to buffering stochastic transcriptional fluctuations in mammalian gene expression.

2. Zusammenfassung

Eine zentrale Frage in der Biologie ist, in welchem Ausmaß zufällige molekulare Prozesse in genetisch identischen Zellen deren deterministische Regulation und Variabilität beeinflussen oder beschränken. Um die Variabilität zwischen einzelnen Zellen zu verstehen, ist es wichtig herauszufinden, wie Zellen ihre Phänotypen stabil an ihre Umgebung und die jeweiligen Wachstumsbedingungen anpassen. Zum Beispiel wird die Zusammensetzung der Plasma Membran an die lokale Zelldichte angepasst, indem das Enzym *focal adhesion kinase* diese erkennt und daraufhin die Transkription bestimmter Gene reguliert. Die Frage, wie sich einzelne menschliche Zellen hinsichtlich der Menge und Lokalisation ihrer Transkripte unterscheiden, stellt den Schwerpunkt meiner Doktorarbeit dar.

Ich habe eine RNA Fluoreszenz in situ Hybridisierungs-Methode entwickelt, mit der ich anhand von hoch auflösenden Bildern einzelne Transkripte in Zellen reproduzierbar quantifizieren kann. Mithilfe neu entwickelter Bild Analyse Verfahren, konnte ich nicht nur die Variabilität von Transkripten zwischen einzelnen Zellen berechnen, sondern auch deren spezifische Lokalisierung im Zytoplasma messen und zeigen, dass diese mit der Funktion des Gens korreliert. Obwohl die Anzahl der Transkripte eines Gens zwischen einzelnen Zellen sehr unterschiedlich ist, ist die Variabilität der meisten Transkripte nur minimal dem Zufall überlassen, und kann sogar anhand von mehrdimensionalen Eigenschaften jeder einzelnen Zelle vorhergesagt werden. Zufällige Fluktuationen in der Genexpression führen zu einer höheren Variabilität in der Anzahl von Transkripten im Kern, jedoch nicht im Zytoplasma. Ich habe hierzu ein computergestütztes Model entwickelt, mit dem ich zeigen konnte, dass die Verweildauer der Transkripte im Zellkern und ihr Export ins Zytoplasma diese Fluktuationen puffert.

3. Introduction

Single isogenic cells exposed to the same conditions show a large degree of cell-to-cell variability (Snijder and Pelkmans, 2011). This variability can be observed from measurements done in single cells for various levels of cellular organization, ranging from fundamental molecular processes to highly complex phenotypic traits. A few examples of these cellular processes and properties that show large cell-to-cell variability are genome organization (Buenrostro et al., 2015; Kind et al., 2013; Nagano et al., 2013) and gene expression (Sigal et al., 2006), the activity of signalling (Feinerman et al., 2008; Spencer et al., 2009) and endocytic pathways (Liberali et al., 2014; Snijder et al., 2009), as well as phenotypes that result from complex molecular interaction, and integrations of different signalling pathways, such as cell cycle timing, cell shape and size, and the propensity of cells to be infected by viruses or to respond to drugs (Bakal et al., 2007; Gut et al., 2015; Snijder et al., 2009; Snijder et al., 2012; Yin et al., 2013).

At the global level, two main sources of cell-to-cell variability have been descried. One source is commonly referred to as intrinsic variability or intrinsic noise, and is the result of stochastic fluctuations on complex chemical reactions in cells, or stochastic partitioning of cellular material during cytokinesis (Elowitz et al., 2002; Paulsson, 2004; Scott et al., 2006; Swain et al., 2002). The second source is generally named extrinsic variability or extrinsic noise, which encompasses heterogeneous types of variability sources, and its definition often depends on the system being studied. For example, the cell cycle stage, the cellular microenvironment, and/or the cell size and volume can be considered sources of extrinsic cell-to-cell variability when studying the variability in endocytosis or virus infection (Elowitz et al., 2002; Paulsson, 2004; Scott et al., 2006; Swain et al., 2002). In contrast, when the source of extrinsic variability is not implicitly defined, it may be estimated as the amount of correlation between otherwise independent processes (Elowitz et al., 2002). In this regard, extrinsic variability can be of stochastic or deterministic nature.

As will be discussed in the latter sections of the introduction, cells have evolved mechanisms to build, regulate and exploit cellular variability.

Since the major part of this thesis focuses on variability of gene expression, I will briefly introduce transcription and other processes related to the life of the mRNA in eukaryotic cells. Finally, it is worth mentioning that although gene expression has been widely studied in the context of cell-to-cell variability, the sources that determine the variability of mRNA levels in the cytoplasm of single isogenic mammalian cells have remained unclear for the vast majority (if not all) genes. Given that transcription is a vital process in the cell, the mechanisms that regulate it and the fundamental limits of such regulation at the single cell level are central questions in biology.

3.1 Regulation and scaling as sources of cell-to-cell variability

Much of the observed cell-to-cell variability is a consequence of biological regulation and/or scaling of cellular activities or components to a given cellular phenotype. Revealing the sources of such regulated variability, often requires a combination of multivariate readouts, advanced statistical analysis, and a series of perturbation experiments (Snijder and Pelkmans, 2011). However, it becomes apparent from the literature that for many cellular systems the overall cell-to-cell variability tends to be dominated by extrinsic sources and arises from tractable biological processes.

A number of studies carried out in the last two decades focused in the structure of variability in protein levels in *E. coli* and yeast. In a seminal study, Elowitz *et al.* used a dual reporter system, in which the expression of YFP and CFP were driven by copies of the same promoter located at equal distance from the origin of replication of the *E. coli* chromosome to tease out the impact of independent stochastic fluctuations at the transcription sites in contrast to upstream shared

sources of variability (Elowitz et al., 2002). They concluded that extrinsic variability contributed to a greater extent to variability in protein levels than intrinsic variability (Elowitz et al., 2002). Using a similar approach, O'Shea and colleagues came to the same conclusion in the yeast *S. cerevisiae,* although the amount of intrinsic variation they observed varied for different genes (Raser and O'Shea, 2004). They showed that the amount of intrinsic variation observed is gene dependent, and argued that gene variability is a trait that can be selected and tuned during evolution (Raser and O'Shea, 2004). Other studies showed that cell size is a major determinant of the variability of protein levels observed in yeast (Newman et al., 2006). Likewise, recent studies in mammalian cells using RNA sm-FISH and single cell RNA-seq demonstrated that cell volume and cell cycle stage greatly impact the levels of transcripts at the single cell level, thus shaping variability (Buettner et al., 2015; Padovan-Merhar et al., 2015).

Together, these data show that the individual levels of gene expression are highly regulated in single cells allowing for scaling of their products to cell cycle stage, cell size and/or volume. Such scaling results in regulated cell-to-cell variability.

Similar regulation at the single cell level has been shown to impact a wide range of cellular activities. A prime example of this is the fact that the cellular microenvironment and the local cell density at which cells grow can be used to predict the pattern of heterogeneity observed in virus infection experiments (Snijder et al., 2009; Snijder et al., 2012). SV40 infection of A431 cells tends to occur primarily in cells located at the periphery of dense regions (Snijder et al., 2009). The observed virus infection heterogeneity was then linked to the different plasma membrane composition of the according host cells (Snijder et al., 2009). Since then, in combination with the cell cycle, the population context and cellular microenvironment has been shown to determine much of the single cell activity of most endocytic pathways (Liberali et al., 2014), as well as signalling by AKT and the ERK pathway and the state of the cytoskeleton (Gut et al., 2015).

3.2 Cells exploit variability

Higher eukaryotic cells have evolved mechanism to generate and exploit cell-to-cell variability. One example is cell development, where stem cells generate and maintain large variation in pluripotency markers, which can be subsequently exploited to trigger different gene expression programs culminating in the formation of a variety of cell types (approximately 10^{14} in mammals) (Arias and Hayward, 2006). The variation of pluripotency markers that has been observed in different stem cell systems is the result of their slow temporal oscillation at the mRNA and protein level. The multipotent mouse haematopoietic cell lime EML, for instance, shows a large variability of the stem cell marker Sca-1 (Chang et al., 2008). When the top 15% of expressing cells are isolated, they take about nine days to relapse to the original distribution. These cells have relatively high levels of PU.1 transcription factor, and are prone to differentiate in the Myeloid lineage. In contrast, the 15% of cells with very low Sca-1 expression levels have relatively high levels of Gata1 and differentiate faster into erythrocytes in response to erythropoietin (Chang et al., 2008). Similarly, neural progenitor cells show a large variation of the transcription factors Asl1, Hes1 and Olig2, which regulate differentiation into neurons, astrocytes and oligodendrocytes, respectively (Imayoshi et al., 2013). Asl1, Hes1 and Olig2 oscillate in neural precursor cells with a period of 175, 150 and 375 min respectively, thus creating the measured cell-to-cell variability. In this case the levels of Asl1 and Hes1 are negatively correlated as a result of Hes1-mediated repression of Asl1 activity (Imayoshi et al., 2013). In addition, Hes1 expression oscillates in phase with another precursor gene, Hes5, indicating that the observed oscillations result from regulated processes rather than from independent stochastic fluctuations. (Imayoshi et al., 2013). Analogue to EML cells, neural progenitor cells gated for given expression levels take three days to return to their initial condition, and high levels of a given progenitor drives differentiation to the corresponding pathway (Imayoshi et al., 2013). Another example are mouse embryonic stem (mES) cells, where fluctuations of Nanog influence the potential of cells to differentiate, with cells exhibiting

low levels of Nanog being more prone to differentiation (Kalmar et al., 2009). A recent study using single cell RNA sequencing of pluripotent cells has shown that pluripotency marker partition into co-regulated modules which can be positively or negatively correlated to Polycomb expression. This study also demonstrated that much of the variability observed for the transcripts of these genes is of a regulated nature (Kumar et al., 2014). Indeed, inhibition of the ERK and GSK3 signalling pathways, or Dicer knockout led to dampening of the pluripotency marker cell-to-cell variability and enhanced self-renewal of stem cells (Kumar et al., 2014).

3.3 Transcription and its regulation

Transcription is the process by which cells express the molecular information encoded in the genome. During transcription a segment of the genome of a cell is copied into many RNA molecules, referred to as messenger RNA (mRNA) for protein coding genes (Lee and Young, 2000). Cells regulate the transcriptional activity of different genes in response to different signals and environmental conditions, such as stress, growth hormones, and paracrine signalling or small chemicals. The regulation of the transcription activity of a gene is rather complex, involving *cis* elements and *trans* regulatory factors, as well as chromatin remodelling (Dynan and Tjian, 1985; Ptashne, 1988; Ptashne and Gann, 1997; Shlyueva et al., 2014). It is this regulatory complexity of transcription that gives metazoan cells a large repertoire of possible states that can be exploited to form different tissues through development.

A simple transcriptional unit in metazoans is shown in Figure 3.1. *Cis* regulatory elements, such as enhancers, insulators, proximal promoter elements, and core promoter elements, are highly organized DNA sequences that orchestrate gene regulation and cell differentiation during development (Wittkopp and Kalay, 2012).

Enhancers were first described as sequences that had the ability to enhance transcriptional activity of a given gene although located distant from the according gene promoter (Banerji et al., 1983; Banerji et al., 1981; Schaffner, 2015). Now, it is clear that enhancers bind transcription factors and are responsible for the recruitment of RNA pol II to the core promoters and its activation (Shlyueva et al., 2014). The interaction of a given enhancer with a given promoter is cell type specific, and the regulation of the core promoter activity by enhancers can be achieved by different means. Enhancers show intrinsic specificity to either core promoters of housekeeping genes or core promoters of developmental regulated genes. (Zabidi et al., 2015). Only developmental enhancers can have altered activity in different cell types, thus defining the cell identity (Zabidi et al., 2015). The specificity of enhancer activity is also influenced by insulator elements, which were discovered as DNA sequences able to disrupt the interaction of the enhancer and the promoter when placed between them (Burgess-Beusse et al., 2002), thus inactivating transcription. There are two main models that describe how insulator elements function. One model is sometimes referred to as the looping model, wherein two insulator sequences interact with one another via insulator binding factors (e.g. CTCCC-binding factor, CTCF), thus changing the 3D conformation of the chromatin and preventing interaction of enhancers with promoters (van Arensbergen et al., 2014). In contrast, the decoy model describes a direct interaction between an insulator and a promoter or an enhancer (van Arensbergen et al., 2014).

The promoter elements are the sequences immediately surrounding the transcription start site (TTS), and are the site of assembly of the pre-initiation complex (Nikolov and Burley, 1997).

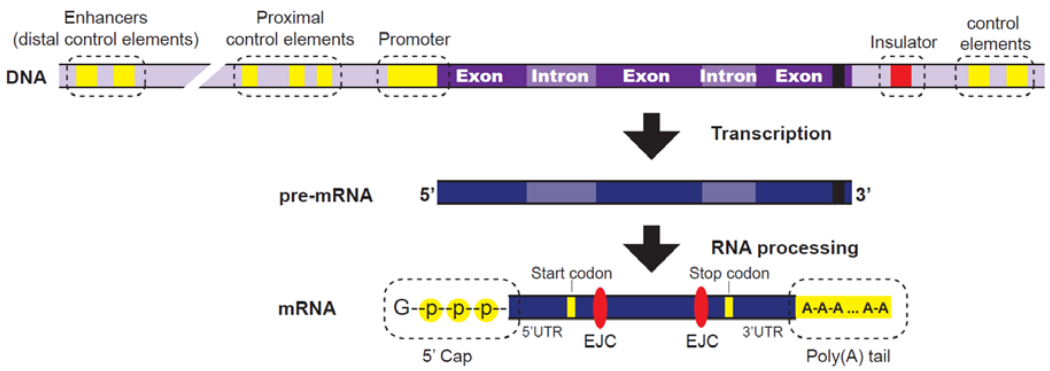


Figure 3.1. A minimal eukaryotic transcriptional unit. The figure shows components of the eukaryotic transcriptional unit and examples of *cis* regulatory elements. EJC stands for 'exon junction complex' and UTR stands for 'untranslated regions'.

Trans regulatory factors are transcription factors that bind specific short sequences (motifs) in the DNA (Spitz and Furlong, 2012). Transcription factors can be activators, and hence increase transcription of a gene, or repressors, which decrease transcription rate (Spitz and Furlong, 2012). Transcriptional activators upon interaction with DNA recruit general transcription factors and the mediator complex, which in turns associates with RNA PolII, to allow the formation of the initiation complex and the start of transcription (Allen and Taatjes, 2015). In addition, transcription can also be regulated by modification of the chromatin structure, e.g. by covalent modification of histones, to enable or disable the assembly of the transcriptional machinery (Spitz and Furlong, 2012). For example, treatment of cells with oestrogen leads to a sequential recruitment of histone acetyl transferases and other cofactors to the site where the transcription factor oestrogen receptor α binds, and culminates with the recruitment of RNA Pol II and the start of transcription (Coulon et al., 2013; Metivier et al., 2003). In contrast, transcriptional either compete with transcriptional activators for DNA motifs, or bind to and sequester transcriptional activators and general transcription factors (Spitz and Furlong, 2012).

3.4 Measured variability and stochastic models of transcription

The levels of all cellular components vary between single isogenic cells and this cell-to-cell variability arises partially from deterministic regulation and partially from stochastic processes. Stochastic models of transcription have a long history in the literature (Paulsson, 2005), but only in the last decade it was possible to accurately estimate variability in gene expression. In this section I focus on studies aiming to understand the variability of transcript levels in eukaryotic cells.

The variability of transcript abundance in cells can be measured using RNA single-molecule fluorescent *in situ* hybridization (sm-FISH, see Chapter 5). Zenklusen et al., used sm-FISH to capture transcript variability of several genes in the yeast *S. cerevisiae* (Zenklusen et al., 2008). They found that the distribution of transcript abundance of some housekeeping genes (*MDN1*, *KAP104*, and *DOA1*) could be explained by a Poisson distribution (Zenklusen et al., 2008). The later can be approximated using a constitutive model of gene expression (Figure 3.2 a) (Zenklusen et al., 2008), where the mRNA is produced at a constant rate k_1 and degraded by a first order reaction with a rate constant γ_1. One important characteristic of Poisson distributions is that the Fano factor - the variance normalized by the mean of the distribution - is equal to one (Raj and van Oudenaarden, 2009). Interestingly, most studies measuring transcript abundance distributions measure Fano factors that are much higher than one (Raj and van Oudenaarden, 2009). In a seminal work, Raj *et al.* measured the variability of transcripts expressed from a *tetO* promoter integrated into the genome of mammalian cells (Raj et al., 2006). The distributions of transcript abundance showed long tails and could not be explained by the constitutive model of gene expression (Raj et al., 2006). They also observed that cells with a high number of mRNAs were likely to have bright clusters of transcript in the nucleus (as visualized by FISH), meaning that these cells were actively transcribing mRNA (Raj et al., 2006). To explain the observed data they proposed a model where transcription occurred in bursts of mRNA synthesis followed by gaps where no or little transcripts are synthesized (Raj et al., 2006), a phenomenon that had

been described previously for protein expression in yeast and bacteria (Blake et al., 2003; Ozbudak et al., 2002). These results led to the two-state model of transcription (Figure 3.2 b), which assumes that the state of any gene changes with a given probability reflecting different chromatin conformations, either transcription permissive or transcription prohibited (Raj et al., 2006). This model was later generalized to allow more than two states of the gene, and it was found that three- or four-stated were required to faithfully reproduce the response of a gene to an induction signal in *S. cerevisiae* (Neuert et al., 2013).

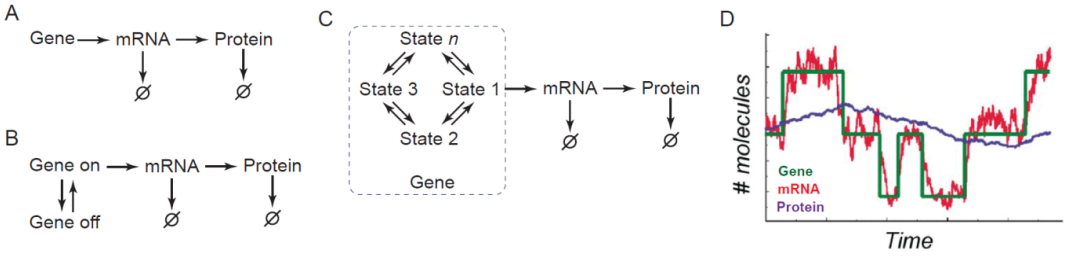


Figure 3.2. Models of stochastic gene expression. A) Constitutive model of mRNA and protein synthesis. **B)** Two-state model of gene expression, in this model the gene state alternates between a silence state where transcription does not occur (off), and a transcriptionally active state (on). **C)** Example of a multistate model of gene expression. In these models the gene can switch between different states reflecting the chromatin conformation, some of which may allow transcription. **D)** Examples of Gillespie's simulations using the two-state model (B). Taken from (Paulsson, 2005).

The dynamics of transcriptional bursts still remains a matter of debate. It is becoming clear that such dynamics can vary greatly depending on the gene and the organism in question. There are two main techniques to measure the dynamics of transcript synthesis in live cells. The first is an indirect method and it relies on the measurement of an unstable protein reporter, either luciferase of a fluorescent protein, to infer the mRNA counts using mathematical models (Molina et al., 2013; Suter et al., 2011) (Figure 3.3). The second technique is to directly measure the activity at the transcription site using the MS2 or PP7 system as shown in Figure 3.3 (Chubb et al., 2006; Larson et al., 2011). Using the indirect reporter system, different groups reported rather slow kinetics of transcriptional bursts, for instance, with an 'on' time ranging from 1 to 20 minutes and 'off' times between 1 to 5 hours, or 'on' time from 1 to 9 hours and 'off' times

from 1 to 14 hours, for different promoter sequences inserted in mammalian cells (Harper et al., 2011; Molina et al., 2013; Suter et al., 2011; Zoller et al., 2015). Because 'off' times of promoters did not follow a single exponential process, these studies concluded that there was a refractory period in the 'off' state before the gene could be switched on again. Such refractory period is interpreted as the time required for the chromatin to reorganize and reopen following the termination of transcriptional events (Suter et al., 2011). This phenomenon cane described with multistate models of chromatin conformation (Figure 3.3D) (Zoller et al., 2015). Only few studies directly measure bursting kinetics at transcription sites using the MS2 systems in mammalian cells, but tend to observe somewhat faster bursting kinetics. For example, the cyclin D1 promoter in HEK 293 cells showed 'on' times of up to 200 minutes and 'off' times between 12 and 36 minutes (Yunger et al., 2010). Likewise, a retroviral construct driven by the EF1α promoter was transcribed with 'on' times of about 2.5 minutes and 'off' times of about 7.5 minutes (Lo et al., 2012). endogenous tagging of the *Nanog* gene with the MS2 system in mouse embryonic stem cells revealed 'on' times ranging from 1.5 to 3 .5 minutes and 'off' times of 11.4 to 35 minutes (Ochiai et al., 2014). Thus, different genes seem to have dramatically different kinetics of transcription.

One important property of such models is that they reproduce all or most experimental distributions of transcript abundance in a population of single cells, and recapitulate the bursts and gaps of transcription observed in live cells for some mammalian promoters. For this reason, it is generally accepted that transcription is a fundamental stochastic process, and intrinsic sources of noise have received more attention in the literature. As a consequence, extrinsic sources of variability are generally not directly modelled in the vast majority of studies.

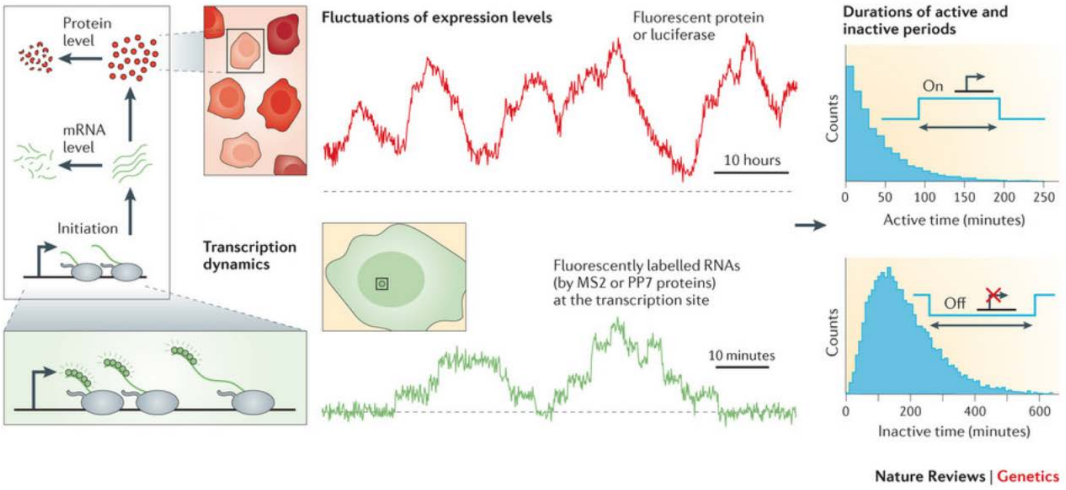


Figure 3.3. The dynamics at the transcription site. Red Traces: Dynamics obtained indirectly by following the fluctuations of a fluorescence reporter or luciferase activity. These constructs generally contain a destabilizing element a the 3'UTR of the transcript to allow accurate recapitulation of synthesis kinetics. Green traces: Direct measurement of the dynamics at a transcription site can be obtained using fluorescently labelled MS2 or PP7 coat proteins targeting the 5'UTR of the transcript. Reproduced from (Coulon et al., 2013).

3.5 The complex life of RNA

The life of RNA molecules from their 'birth'/generation in the nucleus to their (death)/degradation is rather complex. In a nutshell, once the mRNA is transcribed in the nucleus, it undergoes a range of modifications, such as splicing, before it is transported to the cytoplasm, where it can be translated into protein, stored, or degraded. mRNAs are rarely 'naked', but are, immediately after transcription, bound by RNA binding proteins (RBPs) to form messenger ribonucleoparticles (mRNPs). RBPs control much of the life of mRNAs.

Splicing is the process by which introns between the coding sequences (exons) are removed from the mRNA and happens either during transcription or post-transcriptional in the nucleoplasm (Djebali et al., 2012; Tilgner et al., 2012; Vargas et al., 2011). mRNA that has been

spliced and contains a 5' cap is known as mature mRNA, and constitutes the prerequisite for its export from the nucleus (Kohler and Hurt, 2007; Wickramasinghe and Laskey, 2015).

mRNA export through the nuclear pores is a highly regulated step during mRNA biogenesis, and miss-regulation of export or aberrant expression of export machinery elements has been implicated in the development of cancers (Capelson and Hetzer, 2009; Kohler and Hurt, 2010; Siddiqui and Borden, 2012). There are two main export pathways described in eukaryotic cells. The first pathway is mediated by the TAP/p15 receptor and is responsible for the major part of the mRNA export in mammalian cells (Culjkovic-Kraljacic and Borden, 2013; Gruter et al., 1998; Kohler and Hurt, 2007; Reed and Hurt, 2002). Briefly, the transcription export complex (TREX), composed of UAP56 and THO (Kohler and Hurt, 2007), is assembled in the 5'end pre-mRNA molecules during capping and splicing (Kohler and Hurt, 2007; Lei et al., 2001). Aly/REF interactions with the THO complex can assist the loading of the nuclear mRNPs to the TAP/p15 receptor for export (Culjkovic-Kraljacic and Borden, 2013). The second pathway for mRNA export is the CRM1 pathway. CRM1 mainly mediates nuclear export of proteins containing a leucine-rich nuclear export signal (NES) (Dong et al., 2009), thus indirectly mediates export of mRNAs that are bound to proteins containing a NES (Hutten and Kehlenbach, 2007). In addition, the CRM1 pathway is responsible for the export of some mRNAs containing AU-rich elements (AREs) in 3' untranslated regions, e.g. *c-fos* (Gallouzi and Steitz, 2001).

In the cytoplasm, mRNA must be released from the export machinery, which is generally achieved via the ATP-dependent DEAD box helicase DDX19 and the cofactor Gle1 (Folkmann et al., 2011). Newly synthesised and exported transcripts in the cytoplasm are characterized by their association with the cap binding protein (CBP) complex (CBC), CBP80/20, at the 5' end, with the poly-(A) binding protein N1 (PABPN1) and PABPC1 at the 3' end , and with the exon junction complex (EJC) at the exon-exon boundaries (Isken and Maquat, 2008). These mRNAs undergo a quality control, known as "pioneer round" of translation that leads to nonsense mediated decay (NMD) of the mRNA if a stop codon is found upstream of a site bound by an EJC

(Amrani et al., 2006; Isken and Maquat, 2008; Lejeune et al., 2002). In contrast, 'aged' eukaryotic mRNAs synthesized by RNA Pol II are bound at their 5' 7-methylguanosine by the translation initiation factor eIF4E, and at their poly-(A) tail by PABPC1 (Chin et al., 2004). During NMD, CBP80 interacts with the nonsense-mediated decay factor up-frameshift 1 (UPF1) to increase the rate of RNA degradation (Isken et al., 2008).

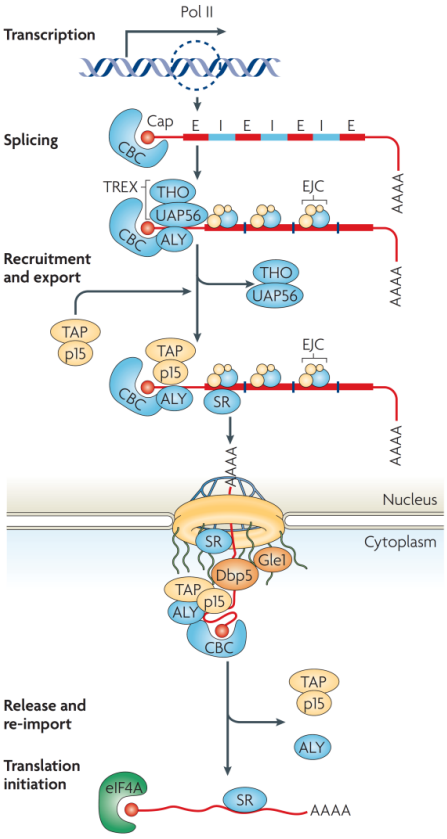


Figure 3.4. mRNA export in metazoans. The main pathway for mRNA export in metazoans is shown. Here TREX recruitment depends on splicing and capping of mRNA. Later recruitment of TAP–p15 mRNA is dependent on the TREX complex. Export is culminated at the cytoplasmic side by the Dbp5/DDX19 helicase and Gle1. EJC stands for 'exon junction complex'. Figure reproduced from (Kohler and Hurt, 2007).

The proteins that bind to the 5'cap and the poly-(A) tail of the mRNA are the main determinants of RNA stability in the cytoplasm, as the main mRNA turnover pathway in eukaryotes involves the sequential shortening of the poly-(A) tail (Norbury, 2013). Poly-(A) tail shortening is carried out by deadenylases, some of which are multiprotein enzymatic complexes. In eukaryotes, the

most studied deadenylases are PAN2/3, which mediate the first round of the poly-(A) tail shortening after transcription, the CCR4-NOT complex, and PARN, which is a 5' cap-dependent deadenylase (Norbury, 2013).

There is evidence that regulation of mRNA degradation involves *cis* acting elements located either distantly from the coding sequence (e.g. the promoter), or at the 3' UTR of the mRNA (e.g. AREs). These sequence elements determine many steps in the life of RNA, such as export, localization in the cytoplasm, and degradation, by promoting association of specific RBPs to the mRNA. In yeast, for example, the promoter responsiveness to heat-shock factor 1 (Hsf1) specifies diffuse cytoplasmic localization of the mRNA upon glucose starvation (Zid and O'Shea, 2014). Likewise, the mRNAs of two mitotic progression regulators SWI5 and CLB2 in yeast are rapidly degraded before mitosis, where the specificity and timing of such degradation determined by their promoter, which directs binding of Dbf2p to the mRNA, which in turns regulates their decay in the cytoplasm (Trcek et al., 2011).

AREs containing transcripts constitute about 10% of the protein encoding transcriptome, and represent highly regulated mRNAs such as, the proto-oncogene *c-fos*, and the inflammatory mediators tumour necrosis factor-α (TNFα), interleukin 1 (IL1), IL2 and IL3 (Khabar, 2005). The stability of ARE containing transcripts is regulated by ARE binding proteins (ARE-BPs) whose stabilizing or destabilizing effects depend on their posttranslational modifications. Examples of ARE-BPs include thetristetrapolin (TTP), butyrate response factor 1 (BRF1), BRF2, and KH-type splicing regulatory protein (KSRP) (Garneau et al., 2007; Maitra et al., 2008), all of which act to destabilize ARE containing transcripts by recruiting exonucleases. In mammalian cells, some components of the mitogen activated protein kinase pathway (e.g. p38 MAPK and MK2), function to regulate the interaction of AREs with ARE-BPs by phosphorylating ARE-BPs and hence modulating their decay (Maitra et al., 2008).

3.6 Aims of the Thesis

In this thesis, I aim to identify mechanisms by which mammalian cells build, control and restrain cell-to-cell variability. One chapter focuses on the modulation of cell-to-cell variability at the membrane by transcriptional regulation, and the remaining chapters of the thesis address variability of transcript abundance and transcript localization in the cytoplasm of single human cells. To this day, the extent of variability in the expression and localization of most endogenous transcripts in mammals is not well known. This is, in part, due to a lack of suitable technologies. Although advances in single cell RNA-seq technologies has been made, they suffer from a low detection efficiency (Grun et al., 2014), which bias measurements of variability (Grun et al., 2014) and they lack spatial resolution. How the several stages of the life of the mRNA impacts the transcript variability introduced during synthesis is also not well understood. Despite a few theoretical studies, experiments addressing this question are lacking. My interest is to increase our understanding of the sources and consequences of such variability, as well as the means by which cells use or constrain variability by: 1) developing high-throughput and computational methods to study endogenous transcripts of mammalian cells *in situ*; 2) identify the sources and determine the structure of variability in transcript abundance as well as define the localization patterns of transcript in single cells; and 3) study what is the consequence on the variability observed of compartmentalizing transcription to the nucleus of cells and, therefore, increasing the complexity of the life of the mRNA.

Chapter 4 describes a detailed mechanism on how variability at the plasma membrane arises from transcriptional control at the single cell level. Briefly, FAK senses the local cell density to regulate transcription of ABCA1, which in turn leads to changes in membrane composition and fluidity. This chapter is a prime illustration of how transcription is regulated in single cells and, thus, determining cell-to-cell variability patterns.

In chapter 5, I describe image-based transcriptomics, a high-throughput experimental methodology developed to study the transcriptome using novel computer vision and machine

learning algorithms. I present a computational method to describe the localization of transcripts within single cells and to analyse the cell-to-cell variability observed in transcript localization. This analysis demonstrates that transcript localization and its variability harbours a higher degree of biological information than transcript abundance. In other words, transcripts of genes with similar biological functions tend to localise with similar patterns in the cytoplasm of cells.

Chapter 6 focuses on the computer vision algorithms that were developed to allow robust detection of nuclei, cells and transcripts required for the image analysis pipelines for image-based trascriptomics. In particular, I present novel algorithms to: 1) perform illumination correction of images exploiting the high number of images acquired; 2) robust detection of nuclei with minimal errors introduced by applying global or local thresholds; and 3) robust detection of cell outlines by iterative application of the Watershed algorithm.

Finally, chapter 7 contains an extensive analysis of the extent and sources of cell-to-cell variability in transcript abundance. Here I show that transcript abundance in the cytoplasm of single human cells is highly variable. Such variability is tightly controlled, so that unaccounted variability achieves a theoretical limit imposed by a single stochastic step for a large number of human genes, and that the latter can be achieved by compartmentalization of transcription in the nucleus. This chapter represents the main analysis of the image-based transcriptomics dataset, as wells as a large number of experiments done in support of the main conclusions.

3.7 References

Allen, B.L., and Taatjes, D.J. (2015). The Mediator complex: a central integrator of transcription. Nat Rev Mol Cell Biol *16*, 155-166.

Amrani, N., Sachs, M.S., and Jacobson, A. (2006). Early nonsense: mRNA decay solves a translational problem. Nat Rev Mol Cell Biol *7*, 415-425.

Arias, A.M., and Hayward, P. (2006). Filtering transcriptional noise during development: concepts and mechanisms. Nat Rev Genet *7*, 34-44.

Bakal, C., Aach, J., Church, G., and Perrimon, N. (2007). Quantitative morphological signatures define local signaling networks regulating cell morphology. Science *316*, 1753-1756.

Banerji, J., Olson, L., and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell *33*, 729-740.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299-308.

Blake, W.J., M, K.A., Cantor, C.R., and Collins, J.J. (2003). Noise in eukaryotic gene expression. Nature *422*, 633-637.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486-490.

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol *33*, 155-160.

Burgess-Beusse, B., Farrell, C., Gaszner, M., Litt, M., Mutskov, V., Recillas-Targa, F., Simpson, M., West, A., and Felsenfeld, G. (2002). The insulation of genes from external enhancers and silencing chromatin. Proc Natl Acad Sci U S A *99 Suppl 4*, 16433-16437.

Capelson, M., and Hetzer, M.W. (2009). The role of nuclear pores in gene regulation, development and disease. EMBO Rep *10*, 697-705.

Chang, H.H., Hemberg, M., Barahona, M., Ingber, D.E., and Huang, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature *453*, 544-547.

Chin, S.Y., Lejeune, F., Ranganathan, A.C., and Maquat, L.E. (2004). The pioneer translation initiation complex is functionally distinct from but structurally overlaps with the steady-state translation initiation complex. Gene Dev *18*, 745-754.

Chubb, J.R., Trcek, T., Shenoy, S.M., and Singer, R.H. (2006). Transcriptional pulsing of a developmental gene. Curr Biol *16*, 1018-1025.

Coulon, A., Chow, C.C., Singer, R.H., and Larson, D.R. (2013). Eukaryotic transcriptional dynamics: from single molecules to cell populations. Nat Rev Genet *14*, 572-584.

Culjkovic-Kraljacic, B., and Borden, K.L. (2013). Aiding and abetting cancer: mRNA export and the nuclear pore. Trends Cell Biol *23*, 328-335.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F.*, et al.* (2012). Landscape of transcription in human cells. Nature *489*, 101-108.

Dong, X., Biswas, A., Suel, K.E., Jackson, L.K., Martinez, R., Gu, H., and Chook, Y.M. (2009). Structural basis for leucine-rich nuclear export signal recognition by CRM1. Nature *458*, 1136-1141.

Dynan, W.S., and Tjian, R. (1985). Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. Nature *316*, 774-778.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183-1186.

Feinerman, O., Veiga, J., Dorfman, J.R., Germain, R.N., and Altan-Bonnet, G. (2008). Variability and robustness in T cell activation from regulated heterogeneity in protein levels. Science *321*, 1081-1084.

Folkmann, A.W., Noble, K.N., Cole, C.N., and Wente, S.R. (2011). Dbp5, Gle1-IP6 and Nup159: a working model for mRNP export. Nucleus *2*, 540-548.

Gallouzi, I.E., and Steitz, J.A. (2001). Delineation of mRNA export pathways by the use of cell-permeable peptides. Science *294*, 1895-1901.

Garneau, N.L., Wilusz, J., and Wilusz, C.J. (2007). The highways and byways of mRNA decay. Nat Rev Mol Cell Bio *8*, 113-126.

Grun, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat Methods *11*, 637-640.

Gruter, P., Tabernero, C., von Kobbe, C., Schmitt, C., Saavedra, C., Bachi, A., Wilm, M., Felber, B.K., and Izaurralde, E. (1998). TAP, the human homolog of Mex67p, mediates CTE-dependent RNA export from the nucleus. Mol Cell *1*, 649-659.

Gut, G., Tadmor, M.D., Pe'er, D., Pelkmans, L., and Liberali, P. (2015). Trajectories of cell-cycle progression from fixed cell populations. Nat Methods *12*, 951-954.

Harper, C.V., Finkenstadt, B., Woodcock, D.J., Friedrichsen, S., Semprini, S., Ashall, L., Spiller, D.G., Mullins, J.J., Rand, D.A., Davis, J.R.*, et al.* (2011). Dynamic analysis of stochastic transcription cycles. PLoS Biol *9*, e1000607.

Hutten, S., and Kehlenbach, R.H. (2007). CRM1-mediated nuclear export: to the pore and beyond. Trends Cell Biol *17*, 193-201.

Imayoshi, I., Isomura, A., Harima, Y., Kawaguchi, K., Kori, H., Miyachi, H., Fujiwara, T., Ishidate, F., and Kageyama, R. (2013). Oscillatory control of factors determining multipotency and fate in mouse neural progenitors. Science *342*, 1203-1208.

Isken, O., Kim, Y.K., Hosoda, N., Mayeur, G.L., Hershey, J.W., and Maquat, L.E. (2008). Upf1 phosphorylation triggers translational repression during nonsense-mediated mRNA decay. Cell *133*, 314-327.

Isken, O., and Maquat, L.E. (2008). The multiple lives of NMD factors: balancing roles in gene and genome regulation. Nat Rev Genet *9*, 699-712.

Kalmar, T., Lim, C., Hayward, P., Munoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J., and Martinez Arias, A. (2009). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. PLoS Biol *7*, e1000149.

Khabar, K.S.A. (2005). The AU-rich transcriptome: More than interferons and cytokines, and its role in disease. J Interf Cytok Res *25*, 1-10.

Kind, J., Pagie, L., Ortabozkoyun, H., Boyle, S., de Vries, S.S., Janssen, H., Amendola, M., Nolen, L.D., Bickmore, W.A., and van Steensel, B. (2013). Single-cell dynamics of genome-nuclear lamina interactions. Cell *153*, 178-192.

Kohler, A., and Hurt, E. (2007). Exporting RNA from the nucleus to the cytoplasm. Nat Rev Mol Cell Biol *8*, 761-773.

Kohler, A., and Hurt, E. (2010). Gene regulation by nucleoporins and links to cancer. Mol Cell *38*, 6-15.

Kumar, R.M., Cahan, P., Shalek, A.K., Satija, R., DaleyKeyser, A.J., Li, H., Zhang, J., Pardee, K., Gennert, D., Trombetta, J.J.*, et al.* (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. Nature *516*, 56-61.

Larson, D.R., Zenklusen, D., Wu, B., Chao, J.A., and Singer, R.H. (2011). Real-time observation of transcription initiation and elongation on an endogenous yeast gene. Science *332*, 475-478.

Lee, T.I., and Young, R.A. (2000). Transcription of eukaryotic protein-coding genes. Annu Rev Genet *34*, 77-137.

Lei, E.P., Krebber, H., and Silver, P.A. (2001). Messenger RNAs are recruited for nuclear export during transcription. Genes Dev *15*, 1771-1782.

Lejeune, F., Ishigaki, Y., Li, X., and Maquat, L.E. (2002). The exon junction complex is detected on CBP80-bound but not eIF4E-bound mRNA in mammalian cells: dynamics of mRNP remodeling. EMBO J *21*, 3536-3545.

Liberali, P., Snijder, B., and Pelkmans, L. (2014). A hierarchical map of regulatory genetic interactions in membrane trafficking. Cell *157*, 1473-1487.

Lo, M.Y., Rival-Gervier, S., Pasceri, P., and Ellis, J. (2012). Rapid transcriptional pulsing dynamics of high expressing retroviral transgenes in embryonic stem cells. PLoS One *7*, e37130.

Maitra, S., Chou, C.F., Luber, C.A., Lee, K.Y., Mann, M., and Chen, C.Y. (2008). The AU-rich element mRNA decay-promoting activity of BRF1 is regulated by mitogen-activated protein kinase-activated protein kinase 2. Rna *14*, 950-959.

Metivier, R., Penot, G., Hubner, M.R., Reid, G., Brand, H., Kos, M., and Gannon, F. (2003). Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. Cell *115*, 751-763.

Molina, N., Suter, D.M., Cannavo, R., Zoller, B., Gotic, I., and Naef, F. (2013). Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. Proc Natl Acad Sci U S A *110*, 20563-20568.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59-64.

Neuert, G., Munsky, B., Tan, R.Z., Teytelman, L., Khammash, M., and van Oudenaarden, A. (2013). Systematic identification of signal-activated stochastic gene regulation. Science *339*, 584-587.

Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441*, 840-846.

Nikolov, D.B., and Burley, S.K. (1997). RNA polymerase II transcription initiation: a structural view. Proc Natl Acad Sci U S A *94*, 15-22.

Norbury, C.J. (2013). Cytoplasmic RNA: a case of the tail wagging the dog. Nat Rev Mol Cell Biol *14*, 643-653.

Ochiai, H., Sugawara, T., Sakuma, T., and Yamamoto, T. (2014). Stochastic promoter activation affects Nanog expression variability in mouse embryonic stem cells. Sci Rep *4*, 7125.

Ozbudak, E.M., Thattai, M., Kurtser, I., Grossman, A.D., and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. Nat Genet *31*, 69-73.

Padovan-Merhar, O., Nair, G.P., Biaesch, A.G., Mayer, A., Scarfone, S., Foley, S.W., Wu, A.R., Churchman, L.S., Singh, A., and Raj, A. (2015). Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. Mol Cell *58*, 339-352.

Paulsson, J. (2004). Summing up the noise in gene networks. Nature *427*, 415-418.

Paulsson, J. (2005). Models of stochastic gene expression. Phys Life Rev *2*, 157-175.

Ptashne, M. (1988). How eukaryotic transcriptional activators work. Nature *335*, 683-689.

Ptashne, M., and Gann, A. (1997). Transcriptional activation by recruitment. Nature *386*, 569-577.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. PLoS Biol *4*, e309.

Raj, A., and van Oudenaarden, A. (2009). Single-molecule approaches to stochastic gene expression. Annu Rev Biophys *38*, 255-270.

Raser, J.M., and O'Shea, E.K. (2004). Control of stochasticity in eukaryotic gene expression. Science *304*, 1811-1814.

Reed, R., and Hurt, E. (2002). A conserved mRNA export machinery coupled to pre-mRNA splicing. Cell *108*, 523-531.

Schaffner, W. (2015). Enhancers, enhancers - from their discovery to today's universe of transcription enhancers. Biol Chem *396*, 311-327.

Scott, M., Ingalls, B., and Kaern, M. (2006). Estimations of intrinsic and extrinsic noise in models of nonlinear genetic networks. Chaos *16*, 026107.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet *15*, 272-286.

Siddiqui, N., and Borden, K.L. (2012). mRNA export and cancer. Wiley Interdiscip Rev RNA *3*, 13-25.

Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Liron, Y., Rosenfeld, N., Danon, T., Perzov, N., and Alon, U. (2006). Variability and memory of protein levels in human cells. Nature *444*, 643-646.

Snijder, B., and Pelkmans, L. (2011). Origins of regulated cell-to-cell variability. Nat Rev Mol Cell Biol *12*, 119-125.

Snijder, B., Sacher, R., Ramo, P., Damm, E.M., Liberali, P., and Pelkmans, L. (2009). Population context determines cell-to-cell variability in endocytosis and virus infection. Nature *461*, 520-523.

Snijder, B., Sacher, R., Ramo, P., Liberali, P., Mench, K., Wolfrum, N., Burleigh, L., Scott, C.C., Verheije, M.H., Mercer, J.*, et al.* (2012). Single-cell analysis of population context advances RNAi screening at multiple levels. Mol Syst Biol *8*, 579.

Spencer, S.L., Gaudet, S., Albeck, J.G., Burke, J.M., and Sorger, P.K. (2009). Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature *459*, 428-432.

Spitz, F., and Furlong, E.E. (2012). Transcription factors: from enhancer binding to developmental control. Nat Rev Genet *13*, 613-626.

Suter, D.M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., and Naef, F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. Science *332*, 472-474.

Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci U S A *99*, 12795-12800.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigo, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res *22*, 1616-1625.

Trcek, T., Larson, D.R., Moldon, A., Query, C.C., and Singer, R.H. (2011). Single-Molecule mRNA Decay Measurements Reveal Promoter-Regulated mRNA Stability in Yeast. Cell *147*, 1484-1497.

van Arensbergen, J., van Steensel, B., and Bussemaker, H.J. (2014). In search of the determinants of enhancer-promoter interaction specificity. Trends Cell Biol *24*, 695-702.

Vargas, D.Y., Shah, K., Batish, M., Levandoski, M., Sinha, S., Marras, S.A., Schedl, P., and Tyagi, S. (2011). Single-molecule imaging of transcriptionally coupled and uncoupled splicing. Cell *147*, 1054-1065.

Wickramasinghe, V.O., and Laskey, R.A. (2015). Control of mammalian gene expression by selective mRNA export. Nat Rev Mol Cell Biol *16*, 431-442.

Wittkopp, P.J., and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet *13*, 59-69.

Yin, Z., Sadok, A., Sailem, H., McCarthy, A., Xia, X., Li, F., Garcia, M.A., Evans, L., Barr, A.R., Perrimon, N.*, et al.* (2013). A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes. Nat Cell Biol *15*, 860-871.

Yunger, S., Rosenfeld, L., Garini, Y., and Shav-Tal, Y. (2010). Single-allele analysis of transcription kinetics in living mammalian cells. Nat Methods *7*, 631-633.

Zabidi, M.A., Arnold, C.D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature *518*, 556-559.

Zenklusen, D., Larson, D.R., and Singer, R.H. (2008). Single-RNA counting reveals alternative modes of gene expression in yeast. Nat Struct Mol Biol *15*, 1263-1271.

Zid, B.M., and O'Shea, E.K. (2014). Promoter sequences direct cytoplasmic localization and translation of mRNAs during starvation in yeast. Nature *514*, 117-121.

Zoller, B., Nicolas, D., Molina, N., and Naef, F. (2015). Structure of silent transcription intervals and noise characteristics of mammalian genes. Mol Syst Biol *11*, 823.

4. Cell-intrinsic adaptation of lipid composition to local crowding drives social behaviour.

By

Mathieu Frechin, Thomas Stoeger, Stephan Daetwyler, Charlotte Gehin, **Nico Battich**, Eva-Maria Damm, Lilli Stergiou, Howard Riezman Lucas Pelkmans.

Published in *Nature,* 02 July 2015.

doi:10.1038/nature14429

The contribution of Nico Battich to this chapter was to design and write the cell tracking software used in the live experiments presented in the paper.

LETTER

doi:10.1038/nature14429

Cell–intrinsic adaptation of lipid composition to local crowding drives social behaviour

Mathieu Frechin[1], Thomas Stoeger[1,2], Stephan Daetwyler[1], Charlotte Gehin[3], Nico Battich[1,2], Eva-Maria Damm[4], Lilli Stergiou[4], Howard Riezman[3] & Lucas Pelkmans[1]

Cells sense the context in which they grow to adapt their phenotype and allow multicellular patterning by mechanisms of autocrine and paracrine signalling[1,2]. However, patterns also form in cell populations exposed to the same signalling molecules and substratum, which often correlate with specific features of the population context of single cells, such as local cell crowding[3]. Here we reveal a cell-intrinsic molecular mechanism that allows multicellular patterning without requiring specific communication between cells. It acts by sensing the local crowding of a single cell through its ability to spread and activate focal adhesion kinase (FAK, also known as PTK2), resulting in adaptation of genes controlling membrane homeostasis. In cells experiencing low crowding, FAK suppresses transcription of the ABC transporter A1 (ABCA1) by inhibiting FOXO3 and TAL1. Agent-based computational modelling and experimental confirmation identified membrane-based signalling and feedback control as crucial for the emergence of population patterns of ABCA1 expression, which adapts membrane lipid composition to cell crowding and affects multiple signalling activities, including the suppression of ABCA1 expression itself. The simple design of this cell-intrinsic system and its broad impact on the signalling state of mammalian single cells suggests a fundamental role for a tunable membrane lipid composition in collective cell behaviour.

Adherent tissue culture cells spread out their cell surface more when experiencing low local crowding than high local crowding, resulting in a higher number of focal adhesions, sites of cellular attachment to the extracellular matrix (ECM), and higher levels of activated FAK (Extended Data Fig. 1a). FAK is recruited to focal adhesions, where it undergoes autophosphorylation, and subsequently recruits and phosphorylates phosphatidylinositol-3-OH kinase (PI(3)K) and many other proteins involved in signalling, cell adhesion and cytoskeletal dynamics[4–6]. FAK may thus, in a cell-intrinsic manner, sense local cell crowding by reacting to the available space and mechanical constraints imposed during cell population growth[7,8], and signal this to downstream cellular functions. To test this, we compared the extent of adaptation of the transcriptome to cellular crowding in adherent embryonic fibroblasts from a FAK-knockout mouse (FAK-KO) with cells from the same background in which FAK was stably re-expressed (FAK-rescue).

A total of 1,014 genes (~5% of the whole genome) adapt their transcript abundance to cellular crowding, of which 80% required the presence of FAK to adapt (Fig. 1a). Although FAK induces genes related to cell growth and proliferation (Extended Data Fig. 1b), it suppresses genes involved in membrane and organelle homeostasis (Fig. 1b) in cells experiencing low crowding, amongst which are 4 ATP-binding cassette (ABC) transporters (*Abca1*, *Abca6*, *Abca9* and *Abcg2*) (Extended Data Fig. 1c). *Abca1* was the overall second most strongly suppressed (~14-fold) gene by FAK (Fig. 1a) and the strongest hit amongst all genes in functional annotation terms related to membrane organization (Fig. 1b). ABC transporters mediate the transport of various substrates across membranes, including phospholipids and cholesterol[9,10].

Single-molecule fluorescence *in situ* hybridization and automated image analysis[3,11] confirmed the transcriptomics results at the single-cell level, showing that FAK controls the abundance of *Abca1* transcripts in single cells to local crowding (Fig. 1c and Extended Data Fig. 1d, e). This adaptation involves low (1–20) and highly variable transcript copy numbers (Extended Data Fig. 1d), and also occurs in the presence of growth factors and cytokines in the medium (Extended Data Fig. 1f).

Predicted candidate transcription factors (see Supplementary Information and Supplementary Table 2) were tested for their involvement in this adaptation using RNA-mediated interference (RNAi) in cells that lack FAK (FAK-KO) and thus highly express *Abca1* independent of crowding. RNAi of *Foxo3*, *Tal1* and *Stat4*, as well as *Lxrb* (liver X receptor beta, also known as *Nr1h2*), the canonical transcription factor driving expression of ABCA1 (ref. 12), reduced *Abca1* transcript abundance in these cells by ~50% (Extended Data Fig. 2a). As TAL1 and FOXO3 are phosphorylated by the serine/threonine kinase AKT, which is activated by PI(3)K downstream of FAK[5], leading to rapid degradation of TAL1 (ref. 13) and inactivation of FOXO3 (ref. 14), we focused on these transcription factors. Chromatin immunoprecipitation (ChIP) experiments (Extended Data Fig. 2b) revealed that in cells lacking FAK, both FOXO3 and TAL1 bind to *Abca1* chromatin independent of cellular crowding. In cells expressing FAK, FOXO3 and TAL1 bind to *Abca1* chromatin at closely located positions only when cells experience high crowding (Fig. 2a). This is in contrast to LXRB, which constitutively binds to *Abca1* chromatin independent of cellular crowding or the presence of FAK (Fig. 2a). Furthermore, western blots of multiple adherent cell lines revealed that cells experiencing low crowding contain higher levels of phosphorylated PI(3)K, AKT and FOXO3 and lower levels of TAL1 than cells experiencing high crowding. Consequently, these cells express a low amount of ABCA1 protein at low cellular crowding. Inhibition of PI(3)K (by wortmannin or LY-294002) lack of FAK (FAK-KO), or inhibition of FAK (by Y15) abolished these differences, leading to ABCA1 expression also in cells experiencing low crowding (Fig. 2b–e and Extended Data Fig. 2c–e). These effects were observed in mouse embryonic fibroblasts, human lung epithelial cells and freshly isolated human keratinocytes. Micropatterns confirmed that cell crowding-dependent expression of ABCA1 stems from the available space of a single cell to adhere to, consistent with a cell-intrinsic mechanism of adaptation (Extended Data Fig. 2f).

To understand if this cell-intrinsic mechanism can drive multicellular pattern formation, we applied single-cell mathematical modelling and computer simulation using a coupled two-level agent-based modelling[15] and differential equation approach (Supplementary Information (mathematical appendix)). The agent-based model simulates the dynamic behaviour of focal adhesions (Supplementary Video 1) and their adhesion potential in multiple single cells of a

[1]Faculty of Sciences, Institute of Molecular Life Sciences, University of Zurich, 8057 Zurich, Switzerland. [2]Life Science Zurich Graduate School, Ph.D. program in Systems Biology. ETH Zurich and University of Zurich, 8057 Zurich, Switzerland. [3]Department of Biochemistry, University of Geneva, 1205 Geneva, Switzerland. [4]Institute of Molecular Systems Biology, ETH Zurich, 8057, Zurich, Switzerland.

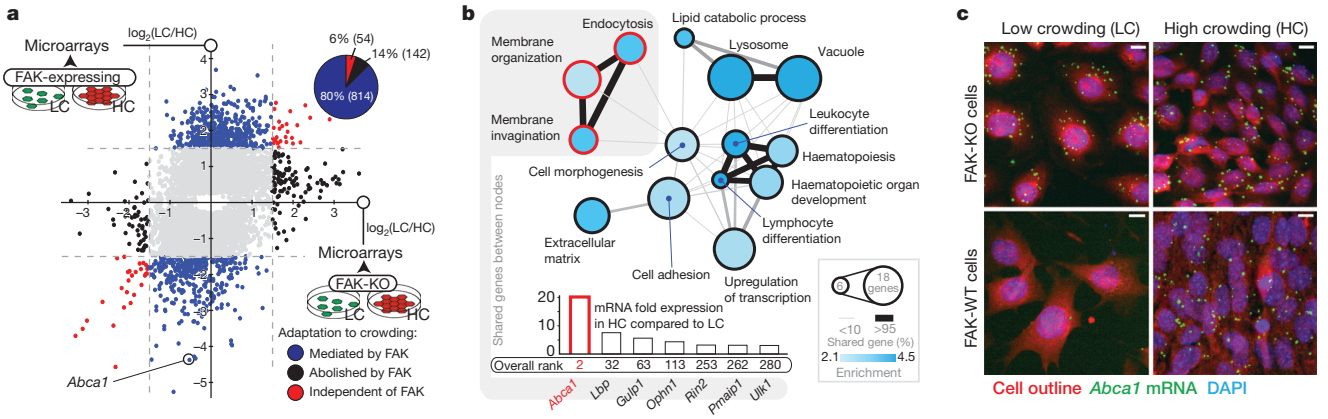


Figure 1 | Adaptation of the transcriptome to cellular crowding. **a**, Scatter plot of the $\log_2$ ratio of transcript abundance in cells experiencing high crowding (HC) over low crowding (LC) in mouse embryonic fibroblasts (MEFs) expressing (FAK-WT, *y* axis) or lacking FAK (FAK-KO, *x* axis). Significance threshold (straight lines): $|\log_2(LC/HC)| > 1.5$. **b**, Gene Ontology enrichment network of genes suppressed by FAK in cells experiencing low crowding. Node colour: enrichment, node size: number of genes, edge width: number of overlapping genes between nodes. **c**, Branched DNA (bDNA) single-molecule FISH against *Abca1* transcripts in FAK-KO (representative of 1.2×10^4 cells) or FAK-WT (representative of 1.5×10^4 cells) MEFs experiencing low or high crowding. DAPI, 4′,6-diamidino-2-phenylindole. Scale bars, 10 μm.

growing cell population (Fig. 3a, Supplementary Video 2 and Supplementary Information). Through indirect constraints that cells impose on each other, emergent properties at both the single-cell and the cell population level arise, including the formation of regions with higher and lower local cell crowding and the emergence of cell polarization and directed migration, agreeing with time-lapse measurements of populations of proliferating cells (Extended Data Fig. 3a–d). In the model, the adhesion potential of each simulated focal adhesion is then used to promote the activation of FAK through an autophosphorylation-based positive feedback loop (Extended Data Fig. 4a). This predicts the appearance of a stable pattern of activated FAK in a population of cells as observed in experiments (Fig. 3b, Supplementary Video 3 and Extended Data Fig. 5a–d).

When modelling suppression of *ABCA1* transcription downstream of activated FAK, we discovered that a gradual pattern of ABCA1 in a growing cell population only emerges when intracellular signal processing is coupled to the timescale at which changes in cellular crowding occur (Fig. 3b and Extended Data Figs 3e and 4b–d), adapted by a feedback mechanism (Fig. 3b, Extended Data Figs 3e and 6a and Supplementary Video 3). Timescale coupling could be achieved by the property of the membrane to act as a storage for phosphatidylinositol-3,4,5-triphosphate (PtdIns(3,4,5)P_3 or PIP_3) production by PI(3)K, while adaptation may be achieved by the ability of ABCA1 to alter physical properties of the membrane[16] leading to a decreased lipid ordering and increased diffusion rate of lipids[17], which affects the probability of AKT activation on the membrane by phosphoinositide-dependent kinase 1 (PDK1)[18,19]. We thus modelled the membrane as a 'tunable capacitor' (Fig. 3b and Extended Data Fig. 3e) that stores PIP_3 and that can be perturbed by ABCA1 in its capacity to activate AKT. This generates a pattern of ABCA1 expression similar to experimental observations that is insensitive to fluctuations in most parameters and primarily depends on the strength of ABCA1 feedback (Extended Data Fig. 6b–d). It also recapitulates the dynamics of ABCA1 down-regulation in scratch assays, when cells at high local crowding suddenly become exposed to free space (Extended Data Fig. 4f, g).

To investigate the existence of ABCA1 feedback on the capacitor function of the membrane, we examined whether the naturally observed crowding-dependent cell-to-cell variability in ABCA1

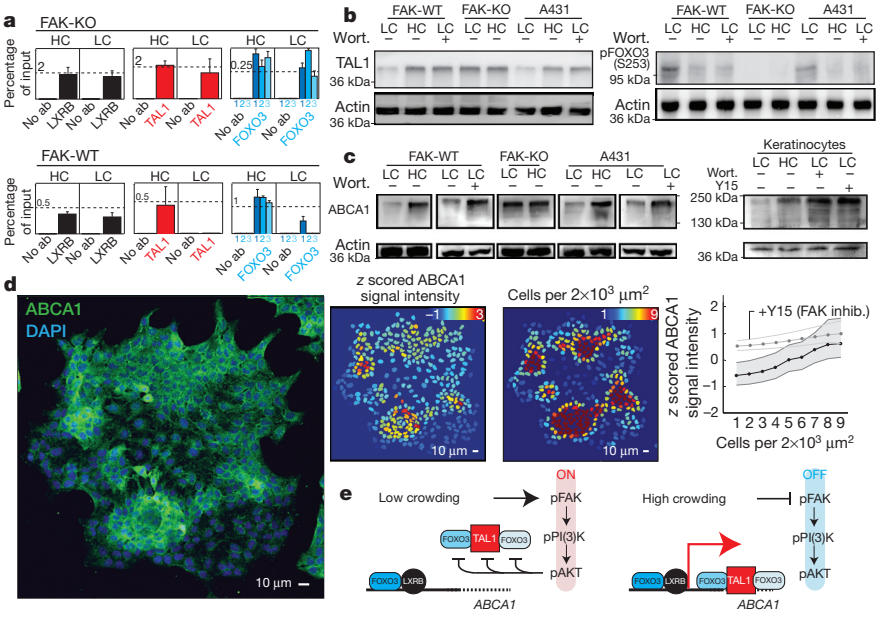


Figure 2 | FAK suppresses ABCA1 expression in cells at low crowding via TAL1 and FOXO3. **a**, ChIP of *Abca1* ($n = 3$ biological replicates, each mean of 3 technical replicates, s.d.) in FAK-KO and FAK-WT MEFs at low or high crowding. No ab, no antibody. **b**, Western blots of pFOXO3 and TAL1 levels in FAK-WT and FAK-KO MEFs, and A431 (human epidermoid carcinoma) cells at low crowding, high crowding, or low crowding + wortmannin. **c**, Western blots of ABCA1 levels as above, including in primary human keratinocytes and upon treatment with Y15 (FAK inhibitor). **d**, Immunofluorescence imaging of ABCA1 in a population of A431 cells (left, representative of 10^4 cells), quantified single-cell intensities of ABCA1 staining (centre) and local cell crowding (right). Curves, single-cell ABCA1 intensities against local cell crowding with or without Y15. Interquartile area in grey, *n* (single cells) > 10^4. **e**, Diagram of the FAK–ABCA1 pathway at low and high crowding. pFAK, phosphorylated FAK.

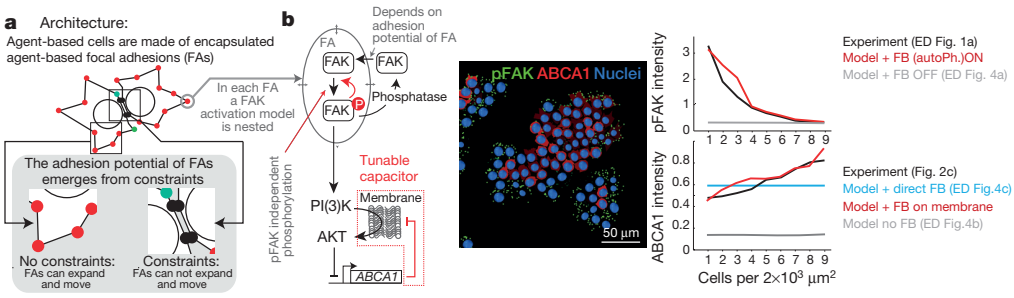


Figure 3 | Multi-scale model of the FAK–ABCA1 system. **a**, Architecture of agent-based modelled single cells encapsulating multiple agent-based modelled focal adhesions. **b**, Model of FAK activation nested in each focal adhesion, influenced by the adhesion potential of each focal adhesion emerging from **a** (left, top part). Model-simulated pFAK levels in single cells (centre image, green signal, representative of all simulations using the same parameters, this run: 10^3 cells) and quantification (right, top graph) against local cell crowding without (grey, Extended Data Fig. 4a) and with (red) positive feedback (FB), experiments in black (Extended Data Fig. 1a). Control of *ABCA1* transcription by FAK using a tunable membrane capacitor topology, which involves PI(3)K and AKT and feedback by ABCA1 (left, bottom part). Model-simulated ABCA1 levels in single cells (centre, red signal), and quantification (right, bottom graph) against local cell crowding without feedback (grey, Extended Data Fig. 4b), with direct feedback (light blue, Extended Data Fig. 4c), and with tunable capacitor (red). Experiments in black.

expression causes changes in membrane lipid composition. Cells experiencing high crowding have a strikingly different lipid composition than cells experiencing low crowding (Fig. 4a and Supplementary Table 3). Cells experiencing low crowding which expressed ABCA1 at levels naturally found in cells experiencing high crowding. (Extended Data Fig. 7a) have a lipid composition more closely resembling that of cells experiencing high crowding (Fig. 4a and Extended Data Fig. 7b). In particular, cells at low crowding have a higher amount of free cholesterol, higher levels of cholesteryl esters (Fig. 4b and Extended Data Fig. 7c), more lipid droplets (Extended Data Fig. 7f), a higher ratio of glucosylceramide over ceramide (GlcCer/Cer) (indicative of glycosphingolipid biosynthesis rate), higher levels of saturated lipids, and lower levels of monounsaturated and polyunsaturated lipids than cells at high crowding (Fig. 4b and Extended Data Fig. 7d, e). In cells experiencing high crowding, plasmid-driven expression of ABCA1 did not alter lipid composition (Fig. 4a, b, and Extended Data Fig. 7b–e). As a consequence, cells experiencing high crowding display lower membrane lipid ordering than cells experiencing low crowding (Fig. 4c), mediated by the crowding-dependent expression of ABCA1 (Extended Data Fig. 7g). Cells that lack FAK and thus express high levels of ABCA1 contain less cholesterol and less of the glycosphingolipid GM1 and display lower membrane lipid ordering than cells expressing FAK (Extended Data Fig. 7h, i).

Similarly, we found that ABCA1 levels influence the amount of S241-phosphorylated PDK1 and T308-phosphorylated AKT (Fig. 4d). Accordingly, levels of T308-phosphorylated AKT are higher in cells experiencing low crowding than cells experiencing high crowding (Fig. 4e). Pharmacological inhibition of ABCA1 abolished this pattern, increasing the level of T308-phosphorylated AKT in cells experiencing high crowding, as predicted by the model when the

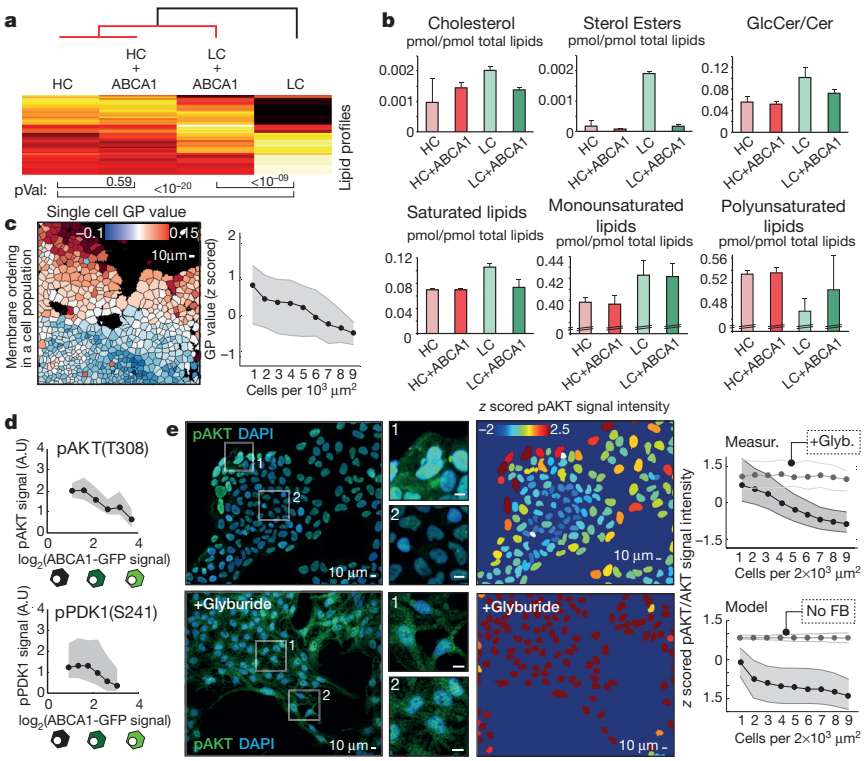


Figure 4 | The FAK–ABCA1 system adapts membrane lipid composition, ordering and signalling to local crowding. **a**, Hierarchical clustering of lipid profiles, see Extended Data Fig. 6b and Supplementary Table 3. *P* values determined by *t*-test. **b**, Histograms of selected lipid species (for free cholesterol in nmol per cell, see Extended Data Fig. 7c). For *P* values (*t*-test), see Extended Data Fig. 7d ($n = 4$ biological replicates, each the mean of 4 technical replicates, s.d.). **c**, *Z*-scored general polarization (GP) values (see Extended Data Fig. 7g) per single A431 cells (left) stained with Laurdan against local cell crowding (right) (interquartile area in grey, number of single cells $> 3 \times 10^3$). **d**, The effect of levels of ABCA1–GFP, randomly expressed from a plasmid in A431 cells at low crowding on pAKT and pPDK1 in single cells (interquartile area in grey). **e**, Untreated (top panels) or glyburide-treated (bottom panels) A431 cells immunostained against pAKT (T308). Nucleus segmentation images are colour-coded for pAKT levels. Top curves (left): single-cell pAKT levels against local crowding in absence (grey) or presence of glyburide (white) (n single cells $> 10^4$). Bottom curves: model-predicted pAKT levels against local crowding with (grey) or without (white) feedback (interquartile areas in grey).

double-negative feedback is removed (Fig. 4e). In addition, exogenous loading of the membrane with cholesterol and the glycosphingolipid GM1, as well as pharmacological inhibition of ABCA1, increases the level of phosphorylated PDK1 and AKT in cells lacking FAK (Extended Data Fig. 7j). Thus, ABCA1 inhibits the FAK-induced signalling pathway that suppresses its own transcription by adapting membrane lipid composition, confirming the membrane-based feedback predicted by the model as a requirement for gradual patterning. We made similar observations for levels of phosphorylated STAT3 and PAK1/2, which are respectively an effector of cytokine receptors and of the small GTPase RAC1, both sensitive to membrane lipid composition (Extended Data Fig. 8)[20,21]. This indicates that the adaptation of membrane lipid composition to local crowding by the FAK–ABCA1 system influences multiple signalling pathways in cells, including those involved in cell motility and paracrine signalling.

We have uncovered a cell-intrinsic molecular mechanism that allows patterning of membrane lipid composition and signalling according to local crowding in a cell population. Several genes with roles in membrane homeostasis may participate in this patterning system, including multiple ABC transporters and lipid-processing enzymes (see Supplementary Table 1, Extended Data Fig. 9 and Supplementary Discussion). In our minimal model, pattern formation of membrane lipid composition only requires variation in the extent of cellular crowding to emerge as cells proliferate. Patterning is subsequently promoted and stabilized by feedback loops without the need for specific cell–cell communication. Because lipid composition affects many membrane protein activities, adapting it to local crowding may have a fundamental role in controlling cellular behaviour within a social context, from colony formation in unicellular organisms[22] to collective cell migration[23], haematopoiesis[24] and T cell activation[25], and the control of epithelial cell proliferation in multicellular organisms[26].

Our work indicates a crucial role for membrane-based signalling in this cell-intrinsic system, in which the membrane may act as a capacitor that converts signals to the correct timescale and is tuned by enzymes that alter membrane lipid composition and ordering in a feedback mechanism. Both timescale adaptation and feedback are required for gradual patterns in a growing cell population to emerge. It will now be important to unravel how such a tunable capacitor operates mechanistically, and to generalize this concept to the possible uses of cellular structures in signal computation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 September 2014; accepted 25 March 2015.
Published online 25 May 2015.

1. Kicheva, A., Cohen, M. & Briscoe, J. Developmental pattern formation: insights from physics and biology. *Science* **338,** 210–212 (2012).
2. Tabata, T. Genetics of morphogen gradients. *Nature Rev. Genet.* **2,** 620–630 (2001).
3. Snijder, B. *et al.* Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature* **461,** 520–523 (2009).
4. Guan, J. L. & Shalloway, D. Regulation of focal adhesion-associated protein tyrosine kinase by both cellular adhesion and oncogenic transformation. *Nature* **358,** 690–692 (1992).
5. Schaller, M. D. Cellular functions of FAK kinases: insight into molecular mechanisms and novel functions. *J. Cell Sci.* **123,** 1007–1013 (2010).
6. Mitra, S. K., Hanson, D. A. & Schlaepfer, D. D. Focal adhesion kinase: in command and control of cell motility. *Nature Rev. Mol. Cell Biol.* **6,** 56–68 (2005).
7. Puliafito, A. *et al.* Collective and single cell behavior in epithelial contact inhibition. *Proc. Natl Acad. Sci. USA* **109,** 739–744 (2012).
8. Piccolo, S., Dupont, S. & Cordenonsi, M. The biology of YAP/TAZ: Hippo signaling and beyond. *Physiol. Rev.* **94,** 1287–1312 (2014).
9. Tarling, E. J., Vallim, T. Q. D. A., Edwards, P. & a.. Role of ABC transporters in lipid transport and human disease. *Trends Endocrinol. Metab.* **24,** 342–350 (2013).
10. Lawn, R., Wade, D. & Garvin, M. The Tangier disease gene product ABC1 controls the cellular apolipoprotein-mediated lipid removal pathway. *J. Clin. Invest.* **104,** 25–31 (1999).
11. Battich, N., Stoeger, T. & Pelkmans, L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nature Methods* **10,** 1127–1133 (2013).
12. Costet, P., Luo, Y., Wang, N. & Tall, A. R. Sterol-dependent transactivation of the ABC1 promoter by the liver X receptor/retinoid X receptor. *J. Biol. Chem.* **275,** 28240–28245 (2000).
13. Palamarchuk, A. *et al.* Akt phosphorylates Tal1 oncoprotein and inhibits its repressor activity. *Cancer Res.* **65,** 4515–4519 (2005).
14. Brunet, A. *et al.* Akt promotes cell survival by phosphorylating and inhibiting a forkhead transcription factor. *Cell* **96,** 857–868 (1999).
15. Holcombe, M. *et al.* Modelling complex biological systems using an agent-based approach. *Integr. Biol. (Camb)* **4,** 53–64 (2012).
16. Zarubica, A. *et al.* Functional implications of the influence of ABCA1 on lipid microenvironment at the plasma membrane: a biophysical study. *FASEB J.* **23,** 1775–1785 (2009).
17. Saffman, P. G. & Delbrück, M. Brownian motion in biological membranes. *Proc. Natl Acad. Sci. USA* **72,** 3111–3113 (1975).
18. Lasserre, R. *et al.* Raft nanodomains contribute to Akt/PKB plasma membrane recruitment and activation. *Nature Chem. Biol.* **4,** 538–547 (2008).
19. Landry, Y. D. *et al.* ATP-binding cassette transporter A1 expression disrupts raft membrane microdomains through its ATPase-related functions. *J. Biol. Chem.* **281,** 36091–36101 (2006).
20. Shah, M., Patel, K., Fried, V. A. & Sehgal, P. B. Interactions of STAT3 with caveolin-1 and heat shock protein 90 in plasma membrane raft and cytosolic complexes: preservation of cytokine signaling during fever. *J. Biol. Chem.* **277,** 45662–45669 (2002).
21. del Pozo, M. A. *et al.* Integrins regulate Rac targeting by internalization of membrane domains. *Science* **303,** 839–842 (2004).
22. Vlamakis, H., Chai, Y., Beauregard, P., Losick, R. & Kolter, R. Sticking together: building a biofilm the *Bacillus subtilis* way. *Nature Rev. Microbiol.* **11,** 157–168 (2013).
23. Friedl, P. & Gilmour, D. Collective cell migration in morphogenesis, regeneration and cancer. *Nature Rev. Mol. Cell Biol.* **10,** 445–457 (2009).
24. Yvan-Charvet, L. *et al.* ATP-binding cassette transporters and HDL suppress hematopoietic stem cell proliferation. *Science* **328,** 1689–1693 (2010).
25. Bensinger, S. J. *et al.* LXR signaling couples sterol metabolism to proliferation in the acquired immune response. *Cell* **134,** 97–111 (2008).
26. Lee, B. H. *et al.* Dysregulation of cholesterol homeostasis in human prostate cancer through loss of ABCA1. *Cancer Res.* **73,** 1211–1218 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank B. Snijder for help with single-cell Laurdan quantification, P. Liberali for help with imaging of cell-to-cell variability, Y. Yakimovich for IT infrastructure support, and all members of the laboratory for discussions and support. M.F. was supported by an EMBO and a Marie Curie (301650) fellowship. E.-M.D. was supported by an Oncosuisse fellowship, L.S. was supported by a Bonizzi Theler fellowship. This work is supported by the University of Zurich and the SystemsX.ch RTD Project LipidX.

Author Contributions L.P. supervised and conceived the project, M.F., T.S., E.-M.D. and L.S. performed experiments, C.G. and H.R. performed lipid mass spectrometry, M.F. and N.B. developed computational image analysis methods, M.F. and L.P. performed data analysis, M.F. and S.D. developed mathematical models, M.F. performed mathematical modelling, L.P. and M.F. wrote the manuscript.

Author Information The microarray data set has been uploaded to the NCBI Gene Expression Omnibus as record GSE43873. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.P. (lucas.pelkmans@imls.uzh.ch).

METHODS

Cell culture. Media and reagents were from GibcoBRL. Wild-type MEFs (FAK-WT), or knockout for FAK (FAK-KO), and A431 cells were purchased from ATCC. Mouse embryonic fibroblasts rescued for FAK (FAK-rescue) were a gift from C. Hauck (University of Konstanz, Germany). E. Reichmann and L. Pontiggia provided keratinocyte primary cells (UZH, Zurich). Standard growth conditions were the following, cells were incubated 3 to 4 days using DMEM containing 10% FBS and 1× glutamine (+135 μg ml^{-1} hygromycinB for the FAK-rescue cells) at 37°C under 5% CO_2. Initial cell number was 2×10^5 to 2.5×10^5 cells for 10-cm dishes 3×10^4 to 5×10^4 cells per well for 12 wells plates containing 13 mm coverslips and 2×10^3 to 2.5×10^3 cells per well for 96-well plates. All our cell lines are tested on a monthly basis for mycoplasma contamination using chemiluminescent assay. The service is independent, centralized for all the UZH and provided at the institute of virology of the UZH. Once the desired population pattern is reached (see video in ref. 3, Snijder *et al.* 2009) cells are serum deprived for approximately 12 h and used for subsequent preparations. Wortmannin (100 nM), Y15 (25 μM), LY-294002 (10 μM) and glyburide (25 μM) treatments were performed over approximately 12 h before preparation. Coverslips were mounted on glass slide using Immu-Mount (Thermo Scientific), a water-based mounting medium.

Plasmid transfection. FAK-WT cells grown in 96-well plates or 10-cm dishes were transfected respectively with 80 ng per well or 4 μg of ABCA1 construct carried in the pEGFP-N1 backbone mixed with 0.2 or 10 μl lipofectamine2000 following the manufacturer's specifications. Homo sapiens *ABCA1* coding sequence was synthesized *de novo* and inserted between SacI and SacII restriction sites. The cloned *ABCA1* sequence corresponds to the full-length consensus coding sequence CCDS6762.1.

Cholesterol and GM1 staining. Cells were quickly washed with successive 1× PBS, 5% delipidated BSA, 1× PBS and fixed for 4 min with 4% PFA. Cholesterol was stained using 0.01 mg ml^{-1} filipin (Sigma) for 20 min, after two washes of 5 min in PBS, surface GM1 was stained using 0.2 μg ml^{-1} cholera toxin subunit B (Alexa Fluor 555 conjugate, Invitrogen) for 10 min.

Laurdan live staining. Cells were grown in ibidi μ-Slide 8 well chambers under standard conditions. Five minutes before acquisition, cells were mounted on the microscope (see microscope section) with environmental control and live stained by addition of 6-dodecanoyl-2-dimethylaminonaphthalene (Laurdan, Molecular Probes) and Draq5 (Cell Signaling) at 5 and 0.5 μM final concentrations directly in the medium. Images were acquired within the next 2 min.

Immunostaining. Unless specified, cells were grown following standard procedures. Fixation was performed with 4%PFA for 10 min, permeabilization with 0.1% Triton X-100 for 10 min, blocking with 1% BSA, 50 mM NH_4Cl for 30 min. Primary and secondary antibodies were diluted in blocking solution, treatments were separated by two 30-min PBS washes. Secondary antibody was applied for 1 h (Alexa Fluor 488 or 568 goat anti rabbit antibody, Invitrogen, 1 μg ml^{-1}). Nuclear staining is performed with 1 μM DAPI for 10 min and cell outlines are visualized with Alexa Fluor 647 carboxylic acid succinimidyl ester (Life Science, 10^{-4} dilution) staining for 10 min. For the pFAK staining, primary antibody was applied for 3 h (rabbit anti-pFAK (Y397) antibody, Cell Signaling no. 3283, 1:200) as well as for ABCA1 (rabbit anti-ABCA1 antibody, Abcam ab7360, 1:500). For pAKT (rabbit anti-pAKT (T308) antibody, Cell Signaling no. 2965, 1:1,000), pPDK1 (rabbit anti-pPDK1 (S241), no. 3061, Cell Signaling, 1:1,000), pSTAT3 (rabbit anti-pSTAT3 (T705) antibody, Cell Signaling no. 9131, 1:500) and pPAK1 (rabbit anti-pPAK1/2 (T423/T402) antibody, Cell Signaling no. 2601, 1:200) staining, primary antibody was applied overnight at 4°C.

mRNA bDNA-FISH experiments. FAK-WT cells were grown following standard conditions in 96-well plates. *Abca1* mRNA bDNA-FISH experiments and image based analysis were performed using the protocol and computational method published by our laboratory[11]. Briefly, cells were fixed, permeabilized, and protease K treated for the *Abca1* mRNA specific probe set to access properly its target sequences. A three-step treatment with successive pre-amplifier, amplifier and fluorescent probes hybridization allows the amplification of the mRNA probe signal and the visualization of single *Abca1* mRNAs. Nuclear staining was performed with 1 μM DAPI for 10 min. Cell outlines were visualized with Alexa Fluor 647 carboxylic acid succinimidyl ester (Life Science) (10^{-4} dilution) staining for 10 min.

Microscopes. Laurdan, filipin and cholera toxin B images were acquired with 40× magnification on a Leica SP5 confocal microscope equipped with a UV laser (λ, 355 nm) in addition to the usual set of visible light lasers, for proper stimulation of Laurdan and filipin. Confocal images of pFAK were acquired on a Zeiss LSM710 microscope with 40× magnification (Zeiss NA1.2, C-apochromat, Korr UV-VIS-IR), GFP-FAK total internal reflection fluorescence (TIRF) video images were acquired on a Nikon visiView microscope with 100× magnification. Immunostainings of ABCA1, pS6, pAKT, pPI(3)K, pSTAT3, pPAK1 and mRNA bDNA-FISH images were acquired on an automated Yokogawa CV7000 spinning disk microscope.

Image analysis. All image analysis was performed using CellProfiler[27] following the same procedure we used in previous publications[3,11,28], with the help of additional MATLAB scripts published previously for the calculation of cellular crowding[3] or written specifically for this study for Laurdan image analysis (see specific section). The general image analysis pipeline was as follows. First, nuclei were detected and segmented based on the DAPI or Draq5 stain using IdentifyPrimaryObjects CellProfiler module. Then, cell boundaries were estimated using nuclear propagation in IdentifySecondaryObjects CellProfiler module. Standard CellProfiler texture, intensity, size and shape features were extracted from nucleus and cell regions. We additionally implemented several image analysis steps for the purpose of detection of out of focus images and for the Support Vector Machine (SVM)-based classification[29] of poorly segmented nuclei.

Membrane ordering analysis. A dedicated CellProfiler module has been developed for this study (the code is available upon request) for defining automatically single-cell generalized polarization (scGP) values after nuclear and cell segmentation. This measurement is based on a previous publication[30] and works as follows: images of cells stained with Laurdan (see specific section above for details) are simultaneously acquired in the 400–460 nm (*I*1) and 470–530 nm (*I*2) wavelength windows after stimulation at 355 nm. The GP value is defined for each pixel following the formula:

$$\mathrm{pxGP} = \frac{I1 - I2}{I1 + I2}$$

The mean GP value of each single cell (scGP value) is then defined by the mean of all pxGP values contained in each segmented cell.

Microarray analysis. High and low crowding FAK-rescue and FAK-KO cells were grown for 24 h in 10-cm dishes, in 10 ml of standard medium (described in the cell culture and preparation section). High crowding cells were seeded at a concentration of 10^6 cells per ml and low crowding cells at 0.4×10^5 cells per ml. RNA preparations were done with the Qiagen RNeasy Mini Kit according to the manufacturer's manual, including the optional column DNase treatment.

The quality of the isolated RNA was determined with a NanoDrop ND 1000 (NanoDrop Technologies, Delaware, USA) and a Bioanalyzer 2100 (Agilent, Waldbronn, Germany). Only the samples with a 260/280 nm ratio between 1.8 and 2.1 and an RNA integrity number (RIN) higher than 8 were further processed. Total RNA samples (100 ng) were reverse-transcribed into double-stranded cDNA in presence of RNA poly-A controls, RNA Spike-In Kit, One-Colour (Agilent product number 5188-5282). The double-stranded cDNAs were *in vitro* transcribed in presence of Cy3-labelled nucleotides using a Low Input Quick Amp Labelling Kit, one-colour (Agilent product number 5190–2305). The Cy3-cDNA was purified using an ARNeasy mini kit, Qiagen (product number 74104 or 74106) and its quality and quantity was determined using NanoDrop ND 1000 and Bioanalyzer 2100. Only cDNA samples with a total cDNA yield higher than 2 μg and a dye incorporation rate between 8 pmol μg^{-1} and 20 pmo μg^{-1} were considered for hybridization.

Cy3-labelled cRNA samples (1.65 μg) were mixed with a Agilent Blocking Solution, subsequently randomly fragmented to 100–200 bp at 65°C with Fragmentation Buffer, and resuspended in Hybridization Buffer using a Gene Expression Hybridization Kit (Agilent product number 5188–5242). Target cRNA Samples (100 μl) were hybridized to Whole Mouse Genome 4×44k OligoMicroarrays (Agilent G4122F) for 17 h at 65°C. Arrays were then washed using Agilent GE Wash Buffers 1 and 2 (Agilent product number 5188–5326), according to the manufacturer's instructions (One-Colour Microarray-Based Gene Expression Analysis Manual, http://www.agilent.com). An Agilent Microarray Scanner (Agilent product number G2565BA) was used to measure the fluorescent intensity emitted by the labelled target. The microarray data set has been uploaded to the NCBI Gene Expression Omnibus as record GSE43873, reorganized and filtered data can be downloaded in the Supplementary Information section (MicroarrayData.xls).

Functional enrichment analysis. The Gene Ontology term enrichment analysis was done with DAVID[31,32] on genes significantly more expressed (absolute $\log_2$(low/high crowding) gene expression value over 1.5) in FAK-expressing cells. Functional groups shown in the two networks have an enrichment value superior than 2 and are composed of at least 5 genes.

Selection of candidate transcription factors. The 19 transcription factors screened in the FAK-KO cells for their potential effect on *Abca1* mRNA expression were selected using a combination of three approaches. (1) Candidates have a binding site in all of the top 10 FAK suppressed genes defined with the microarray data. To perform this comparison, we used the Pscan algorithm (http://www.beaconlab.it/pscan) with the JASPAR database[33] (http://jaspar.genereg.net/). (2) Transcription factors having the strongest GO enrichment for lipid

homeostasis or (3) having a reported ChIP binding site or an effect on expression for ABCA1 in the literature (Supplementary Table 2).

siRNA experiments. All siRNAs were purchased from Qiagen. FAK-KO cells were cultured in 24-well plates, using standard conditions until reaching approximately 60% confluency (48–60 h) and transfected by forward transfection. Per well, 25 pmol samples of siRNA were mixed in 25 µl of Opti-MEM and 0.5 µl of Lipofectamine RNAiMAX were mixed with 24.5 µl of Opti-MEM. After 5 min of incubation, solutions were mixed together and incubated for another 20 min at room temperature and transferred on the cultured cells for 60 h before RNA preparation.

qPCR screening. Silenced FAK-KO cells were washed with 1× PBS, RNA samples were prepared using NucleoSpinRNAII kit (Macherey Nagel), cDNA synthesis was carried out with the Transcriptor High Fidelity cDNA Synthesis Kit (Roche) using poly-dT primers, in both cases following the manufacturer's protocol. Quantitative real-time PCR was performed in 384-well plates in an AB7900HT qPCR device (Applied Biosystems) using the following primers, forward ABCA1: 5′-CTGTAGACCTGGAGAGAAGCTTTC-3′, reverse ABCA1: 5′-CAGCTCCA TGGACTTGTTGATGAG-3′ allowing amplification over the twelfth and thirteenth exons contained in all *ABCA1* mRNA variants, and forward GAPDH: 5′-TCAAGGCTGAGAACGGGAAGCTTG-3′, reverse GAPDH: 5′-AGCCTTCT CCATGGTGGTGAAGAC-3′. Relative mRNA amounts were calculated using GAPDH as an internal reference.

Western blotting. A431, FAK-WT and FAK-KO cells were cultured using standard conditions in 10-cm dishes. Low crowding cells were stopped after 2 to 2.5 days of growth, whereas high crowding cells were grown for 6 days (both including a final 12 h of serum starvation). Cells were washed with 1 × PBS and disrupted in lysis buffer (0.5% sodium deoxycholate, 150 mM NaCl, 50 mM Tris-HCl, pH 7.2, 0.1% SDS, 1% Triton X-100, 0.2% NaN_3), and 15 µg of each protein extract was separated using 10% PAGE except for ABCA1 western blotting where 50 µg of protein and 8% PAGE were used. Separated proteins were then transferred onto a membrane (Immobilon-P, 0.45 µm, Millipore) using the humid chamber method. Transfer conditions are 80 mA overnight for ABCA1 western blotting, 250 mA for 90 min otherwise. Membranes were blocked with 4% BSA proteins in 1× TBS-T (1× TBS, 0.1% Tween) for 1 h. Primary antibodies rabbit anti-pFAK (Cell Signaling no. 3283), rabbit anti-pPI(3)K (rabbit anti-pPI(3)K p85/p55 (T458/T199) antibody, Cell Signaling no. 4228), rabbit anti-pAKT ((T308) Cell Signaling no. 2965) were diluted at 1:1,000 and rabbit anti-actin (Cell Signaling no. 8456) at 1:5,000. Rabbit anti-TAL1 (Sc-12984, Santa Cruz) and rabbit anti-pFOXO3 (S253, no. 9466, Cell Signaling) were diluted at 1:200 and rabbit anti-ABCA1 (Abcam ab7360) at 1:500 in blocking buffer. HRP-conjugated secondary anti-mouse (no. 170-6516, BioRad) and anti-rabbit (no. 170-6515, BioRad) antibodies were diluted at 1:5,000 in the same buffer. Primary and secondary antibodies were applied overnight at 4°C and 60 min at room temperature, respectively. Signal was revealed with HRP substrate solution and imaged with a CCD camera (for antibody references see immunostaining section).

ChIP experiments. FAK-KO and FAK-WT cells were cultured using standard conditions in 10-cm dishes. Low crowding cells were stopped after 2 to 2.5 days of growth, whereas high crowding cells were grown for 6 days (both including a final 12 h of serum starvation). Experiments were carried out using the Chromatin Immunoprecipitation (ChIP) Assay Kit from Millipore following manufacturer's specifications except for the following changes. Fixation of cells was performed with 1.6 mM Di-thio bis-succinimidyl propionate (DSP) for 20 min, two short washes with 1× PBS at room temperature, and finally 1% paraformaldehyde for 20 min. 20 µg of anti-TAL1 (Sc-12984, Santa Cruz), anti-FOXO3 (07–702, Millipore) and anti-LXR beta (Sc-34341, Santa Cruz) primary antibodies was added for 15 h at 4°C to the pre-cleared supernatant. Protein A beads were then added for 4 h. Reversion of crosslinking was done for 12 h at 55°C.

Lipid mass spectrometry

Chemicals and lipid standards. DLPC 12:0/12:0 (850335), PE 17:0/14:1 (PE31:1, LM-1104), PI 17:0/14:1 (PI31:1, LM-1504), PS 17:0/14:1 (PS31:1, LM-1304), C17:0 ceramide (860517), C12:0 SM (860583) and Glucosyl C8:0 Cer (860540) were used as internal lipid standards and were purchased from Avanti Polar Lipids Inc. (Alabaster, AL). Ergosterol was used as sterol standard and was purchased from Fluka (Buchs, Switzerland). Methyl tert-butyl ether (MTBE) was from Fluka (Buchs). Methyl amine (33% in absolute ethanol) was from Sigma Aldrich (Steinheim, Germany). HPLC-grade chloroform was purchased from Acros (Geel, Belgium), liquid chromatography–mass spectrometry (LC–MS) grade methanol and LC–MS grade ammonium acetate were from Fluka. LC–MS grade water was purchased from Biosolve.

Cell culture. FAK-WT cells were cultured using standard conditions in 10-cm dishes. Low crowding cells were stopped after 2.5–3 days of growth while high crowding cells were grown for 6 days (both including a final 12 h of serum starvation). Cells were transfected with a human ABCA1-containing plasmid as described above or subjected to the transfection procedure without plasmid after one day of culture for low crowding cells or four days of culture for high crowding cells. Cells facing low or high crowding were collected two days after transfection. Cells were shortly washed with successively 1× PBS, 5% delipidated BSA, and three times with cold 1× PBS, scraped and pelleted at 800*g* for 5 min before lipid extraction.

Lipid analysis. Lipid extracts of 4 biological replicates of each of the 4 conditions (high crowding; high crowding + ABCA1; low crowding; low crowding + ABCA1) were prepared using the MTBE protocol[34] and measurements were made in 4 technical replicates, amounting to a total of 64 measurements at each mass spectrometer. Cell pellets were resuspended into 100 µl of water and transferred into a 2 ml Eppendorf tube. Then 360 µl methanol and a mix of internal standards were added (400 pmol DLPC, 1,000 pmol PE31:1, 1,000 pmol PI31:1, 3,300 pmol PS31:1, 2,500 pmol C12SM, 500 pmol C17Cer and 100 pmol C8GC). Samples were vortexed and 1.2 ml of MTBE was added. Samples were placed for 10 min on a multitube vortexer at 4°C (Lab-tek International) followed by an incubation for 1 h at room temperature on a shaker. Phase separation was induced by addition of 200 µl MS-grade water. After 10 min of incubation at room temperature, samples were centrifuged at 1,000*g* for 10 min. The upper (organic) phase was transferred into a 13 mm glass tube with a Teflon-lined cap and the lower phase was re-extracted with 400 µl artificial upper phase (MTBE/methanol/H_2O 10:3:1.5). In total, 1,500 µl of organic phase was recovered from each samples, split into three parts and dried in a CentriVap Vacuum Concentrator (Labconco). One part was treated by alkaline hydrolysis to enrich for sphingolipids and the other two aliquots were used for glycerophospholipid/phosphorus assay and sterol analysis, respectively. Glycerophospholipids were deacylated according to the method by Clarke & Dawson[35]. Briefly, 1 ml freshly prepared monomethylamine reagent (methylamine/H_2O/n-butanol/methanol at 5:3:1:4 (vol/vol)) was added to the dried lipid extract and then incubated at 53°C for 1 h in a water bath. Lipids were cooled to room temperature and then dried. For desalting, the dried lipid extract was resuspended in 300 µl water-saturated n-butanol and then extracted with 150 µl H_2O. The organic phase was collected, and the aqueous phase was reextracted twice with 300 µl water-saturated n-butanol. The organic phases were pooled and dried in a CentriVap Vacuum Concentrator.

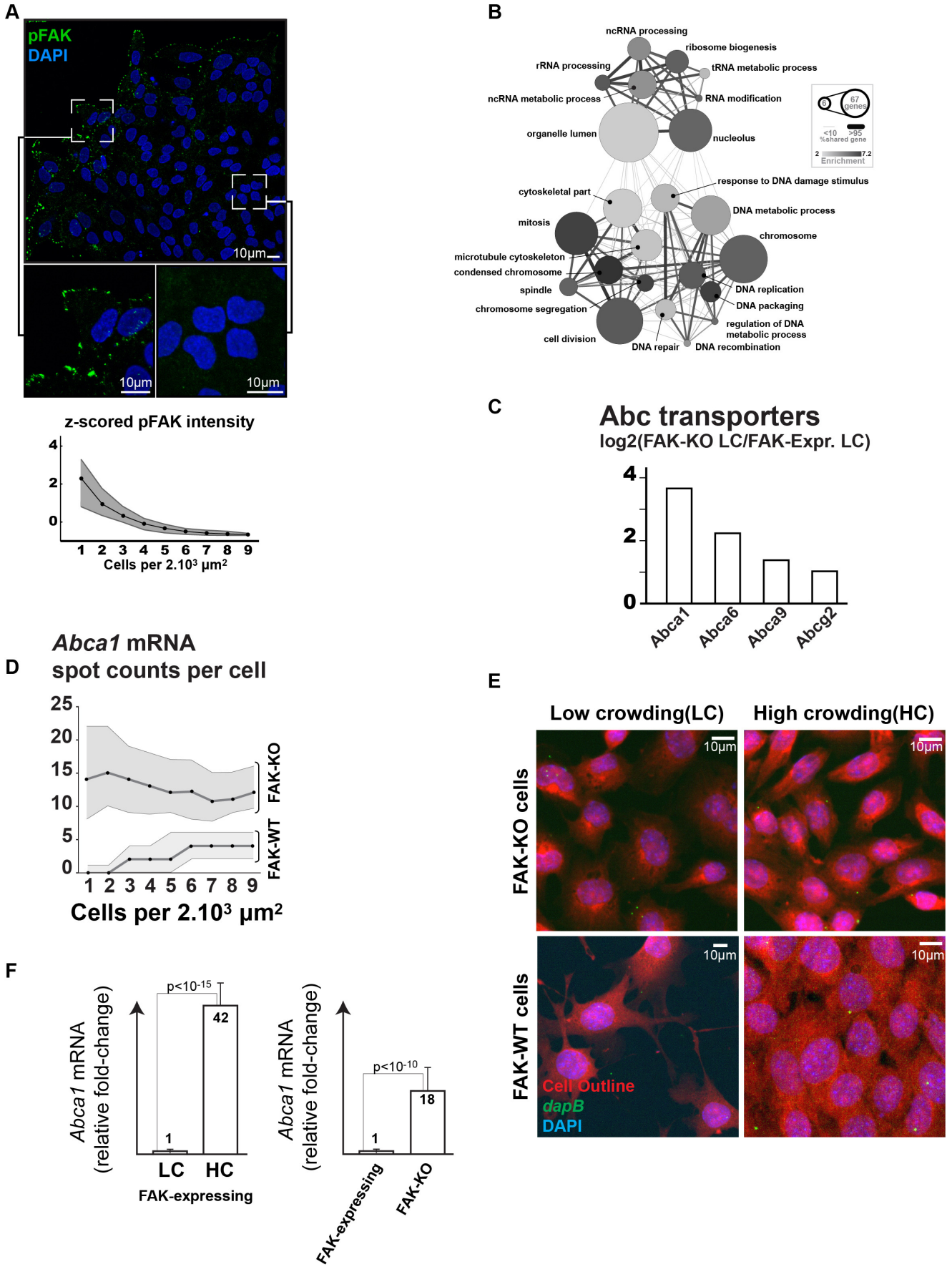
Sterols analysis by gas chromatography–mass spectrometry (GC–MS). One-third of total lipid extract was resuspended in 500 µl of MS-grade chloroform/methanol (1:1) solution and injected into a VARIAN CP-3800 gas chromatogram equipped with a Factor Four Capillary Column VF-5ms 15 mm × 0.32 mm i.d. DF = 100. Identification and quantification of sterol species were performed using a VARIAN 320MS as described in ref. 36.

Phospholipids and sphingolipids analysis by electrospray ionization mass spectrometry (ESI-MS). Identification and quantification of phospholipid and sphingolipid molecular species were performed using multiple reaction monitoring with a TSQ Vantage Triple Stage Quadrupole Mass Spectrometer (Thermo Scientific) equipped with a robotic nanoflow ion source, Nanomate HD (Advion Biosciences). Each individual ion dissociation pathway was optimized with regard to collision energy. Lipid concentrations were calculated relative to the relevant internal standards as described in ref. 37 and then normalized to the total phosphorus content of each total lipid extract to adjust for difference in cell size, membrane content, and extraction efficiency.

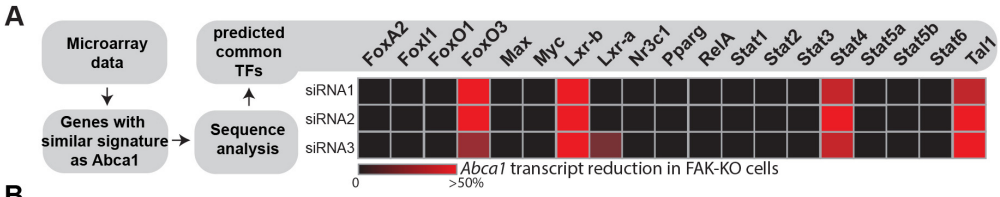
Determination of total phosphorus content. The dried total lipid extract was resuspended in 250 µl chloroform/methanol (1:1) and 50 µl were placed into a 13 mm disposable pyrex tube. The solvent was completely evaporated and 0, 2, 5, 10, 20 µl of a 3 mM KH_2PO_4 standard solution were placed into separate pyrex tubes. To each tube 20 µl of water and 140 µl of 70% perchloric acid were added. Samples were heated at 180°C for 1 h in a hood. Tubes were then removed from the block and kept at room temperature for 5 min. Then 800 µl of freshly prepared H_2O/1.25% NH_4Molybdate (100 mg/8 ml H_2O)/10% ascorbic acid (100 mg/6 ml H_2O) in the ratio of 5:2:1 were added. Tubes were heated at 100°C for 5 min with a marble on each tube to prevent evaporation. Tubes were cooled at room temperature for 5 min. 100 µl of each sample was then transferred into a 96-well microplate and the absorbance at 820 nm was measured[38].

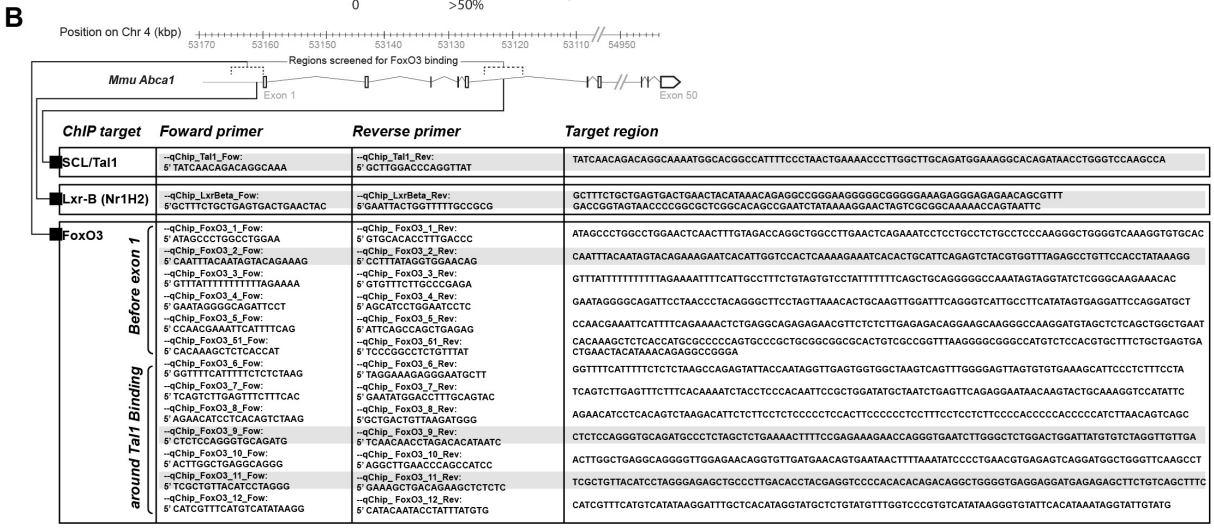
27. Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7,** R100 (2006).
28. Wippich, F. *et al.* Dual specificity kinase DYRK3 couples stress granule condensation/dissolution to mTORC1 signaling. *Cell* **152,** 791–805 (2013).
29. Rämö, P., Sacher, R., Snijder, B., Begemann, B. & Pelkmans, L. CellClassifier: supervised learning of cellular phenotypes. *Bioinformatics* **25,** 3028–3030 (2009).
30. Gaus, K., Zech, T. & Harder, T. Visualizing membrane microdomains by Laurdan 2-photon microscopy. *Mol. Membr. Biol.* **23,** 41–48 (2006).

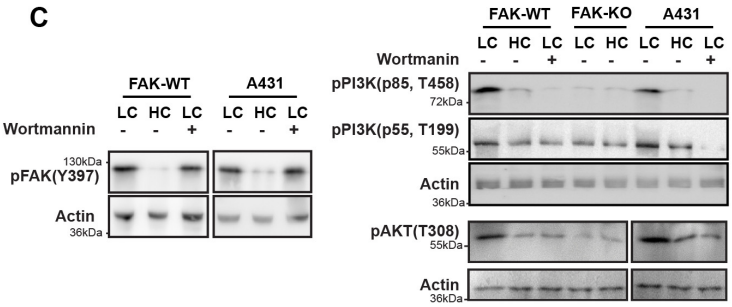
31. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2009).
32. Huang, d. W., Sherman, B. T., Lempicki, R. & a.. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37,** 1–13 (2009).
33. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32,** D91–D94 (2004).
34. Matyash, V., Liebisch, G., Kurzchalia, T. V., Shevchenko, A. & Schwudke, D. Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. *J. Lipid Res.* **49,** 1137–1146 (2008).
35. Clarke, N. G. & Dawson, R. M. *Alkaline O→N-transacylation.* 195**,** 301–306 (1981). A new method for the quantitative deacylation of phospholipids. *Biochem. J.* **195,** 301–306 (1981).
36. Guan, X. L., Riezman, I., Wenk, M. R. & Riezman, H. Yeast lipid analysis and quantification by mass spectrometry. *Methods Enzymol.* **470,** 369–391 (2010).
37. Epstein, S. *et al.* Activation of the unfolded protein response pathway causes ceramide accumulation in yeast and INS-1E insulinoma cells. *J. Lipid Res.* **53,** 412–420 (2012).
38. Rouser, G., Fleischer, S. & Yamamoto, A. Two dimensional thin layer chromatographic separation of polar lipids and determination of phospholipids by phosphorus analysis of spots. *Lipids* **5,** 494–496 (1970).

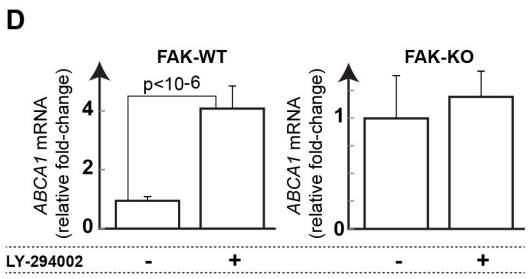


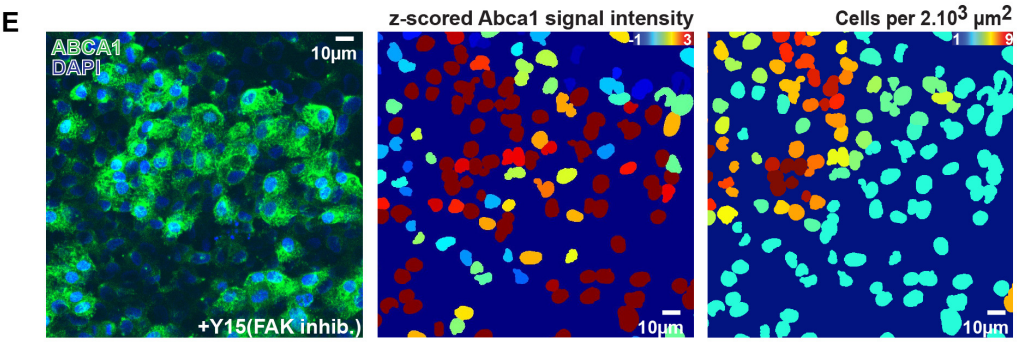
Extended Data Figure 1 | Adaptation of the transcriptome to cellular crowding. Related to Fig. 1. **a**, Immunofluorescence against phosphorylated FAK (Y397) in a population of A431 cells, corresponding curve shows single-cell phosphorylated FAK signals against local cell crowding (interquartile area is shown in grey, number of cells $>10^4$). **b**, Gene Ontology enrichment network of genes that are induced by FAK in cells experiencing low crowding. Greyscale indicates enrichment, node-size number of genes, edge width between nodes number of overlapping genes. **c**, Histogram of ABC transporters more expressed in cells lacking FAK compared to cells expressing FAK when facing low crowding. **d**, Single-cell transcript counts of *Abca1* in 1.2×10^4 FAK-KO and 1.5×10^4 FAK-WT cells experiencing increasing levels of local crowding (interquartile area in grey). **e**, Control experiment of bDNA single-molecule FISH against bacterial *dapB* transcripts in FAK-KO or FAK-WT cells experiencing low crowding or high crowding. Representative of 10^4 cells. **f**, Real-time PCR measurements of *Abca1* transcripts in cells at low and high local crowding in both FAK-expressing and FAK-KO cells in the presence of 10% FCS. Clearly, *Abca1* mRNA levels are much higher in FAK-expressing cells facing high crowding than in the same cells facing low crowding (s.d., $n = 4$ biological replicates each made of 3 technical replicates, $P < 10^{-15}$, *t*-test) but also in FAK-KO cells compared FAK-expressing cells (s.d., $n = 4$ biological replicates each made of 3 technical replicates, $P < 10^{-10}$, *t*-test). This indicates that FAK-dependent adaptation of *Abca1* transcription to cell crowding also operates in the presence of an abundant and homogeneous amount of growth factors and cytokines in the medium.

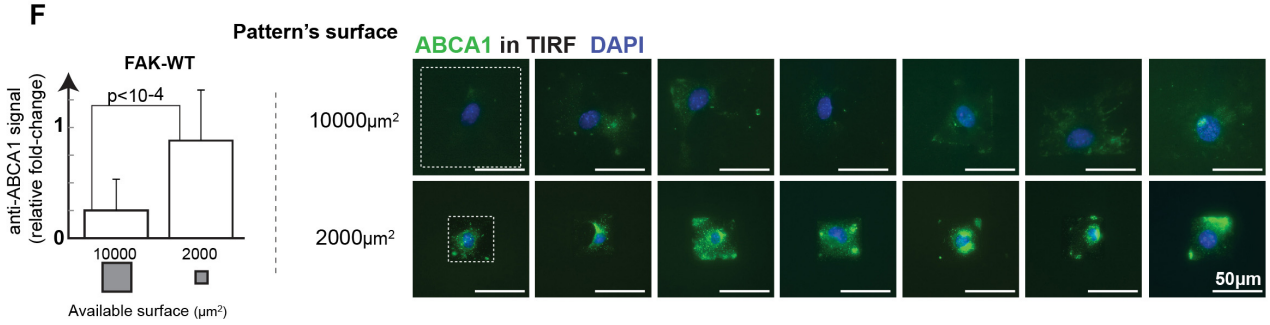




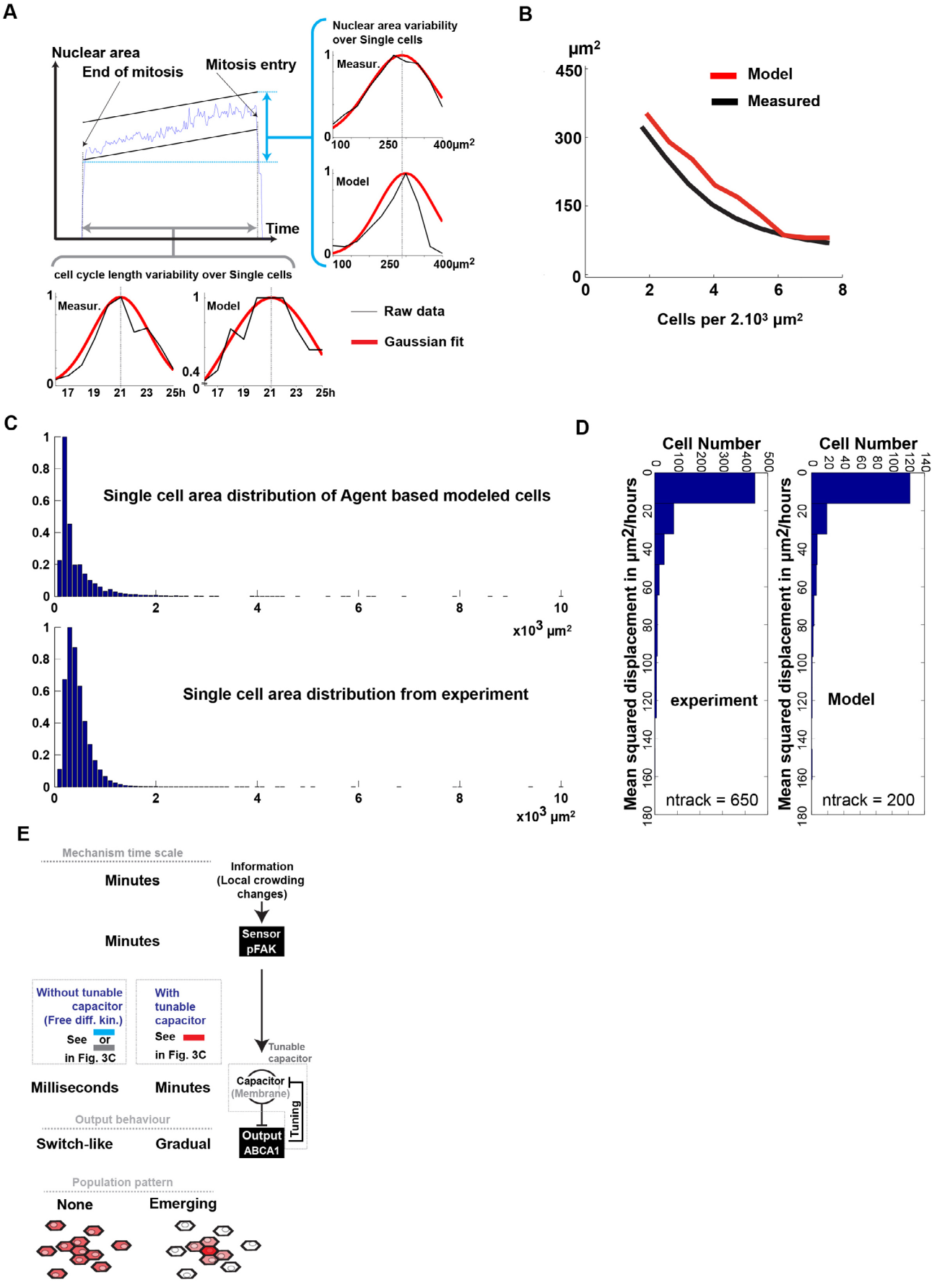




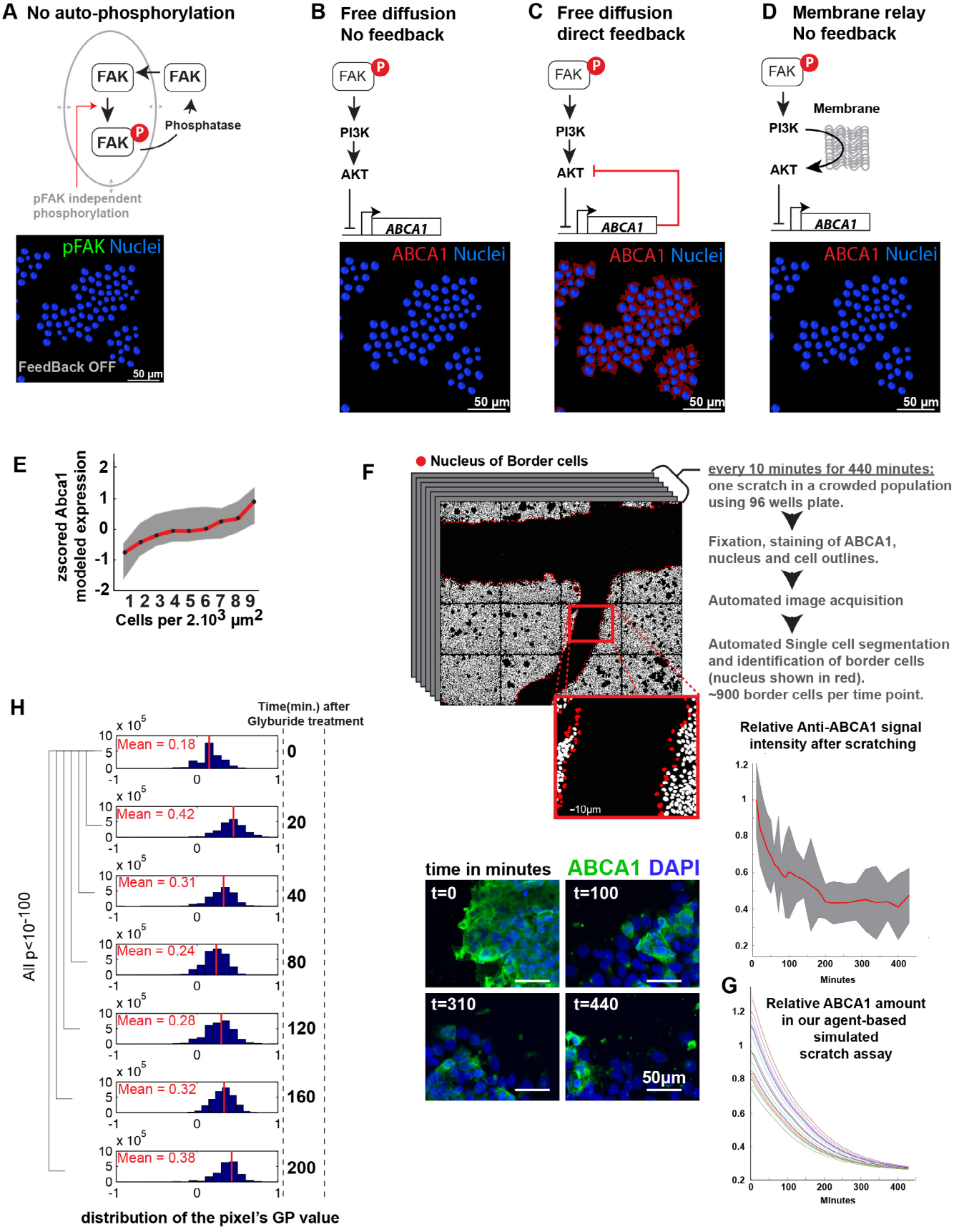




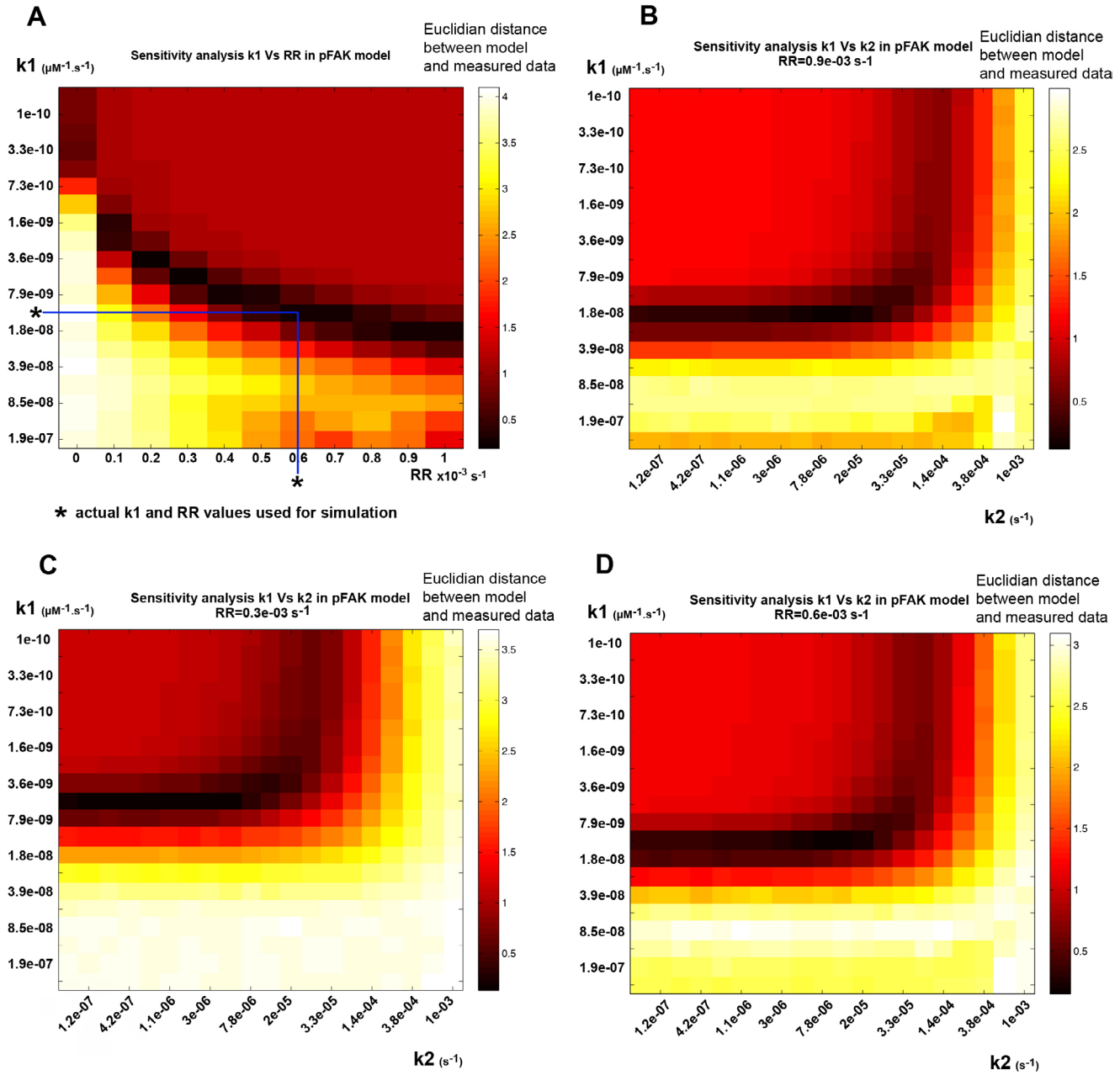
Extended Data Figure 2 | FAK suppresses ABCA1 expression in cells at low crowding via TAL1 and FOXO3 in a cell-intrinsic way. Related to Fig. 2. **a**, Percentage reduction of *Abca1* mRNA in FAK-KO cells upon silencing of 19 potential transcription factors. **b**, Table of primers used for qRT–PCR amplification of *Abca1* DNA and corresponding genomic position. **c**, Western blots of pFAK, pPI(3)K and pAKT levels in FAK-WT and FAK-KO MEFs, and A431 cells at low crowding, high crowding or low crowding + wortmannin. **d**, Real-time PCR quantification of *Abca1* mRNA shows that treatment with LY-294002 alleviates the inhibitory effect of FAK on *Abca1* transcription in cells (at low crowding) expressing FAK (s.d., $n = 4$ biological replicates each made of 3 technical replicates, $P < 10^{-6}$, *t*-test), whereas this treatment has no significant effect on *Abca1* transcription in cells that lack FAK (s.d., $n = 4$ biological replicates each made of 3 technical replicates, $P > 0.1$, *t*-test). **e**, Immunofluorescence imaging of ABCA1 over a population of A431 cells in the presence of Y15 FAK inhibitor and related projection of single cell measurements onto nuclear segmentations. **f**, Quantifications of Abca1 protein expression in FAK-WT cells adhering to micropatterned surfaces of large (10,000 μm^2) or small (2,000 μm^2) area (http://www.cytoo.com) at long distance from potentially secreting neighbouring cells. This shows that space constraints are sufficient to trigger differences in Abca1 expression (s.d., $n = 100$ cells, $P < 10^{-4}$, *t*-test).



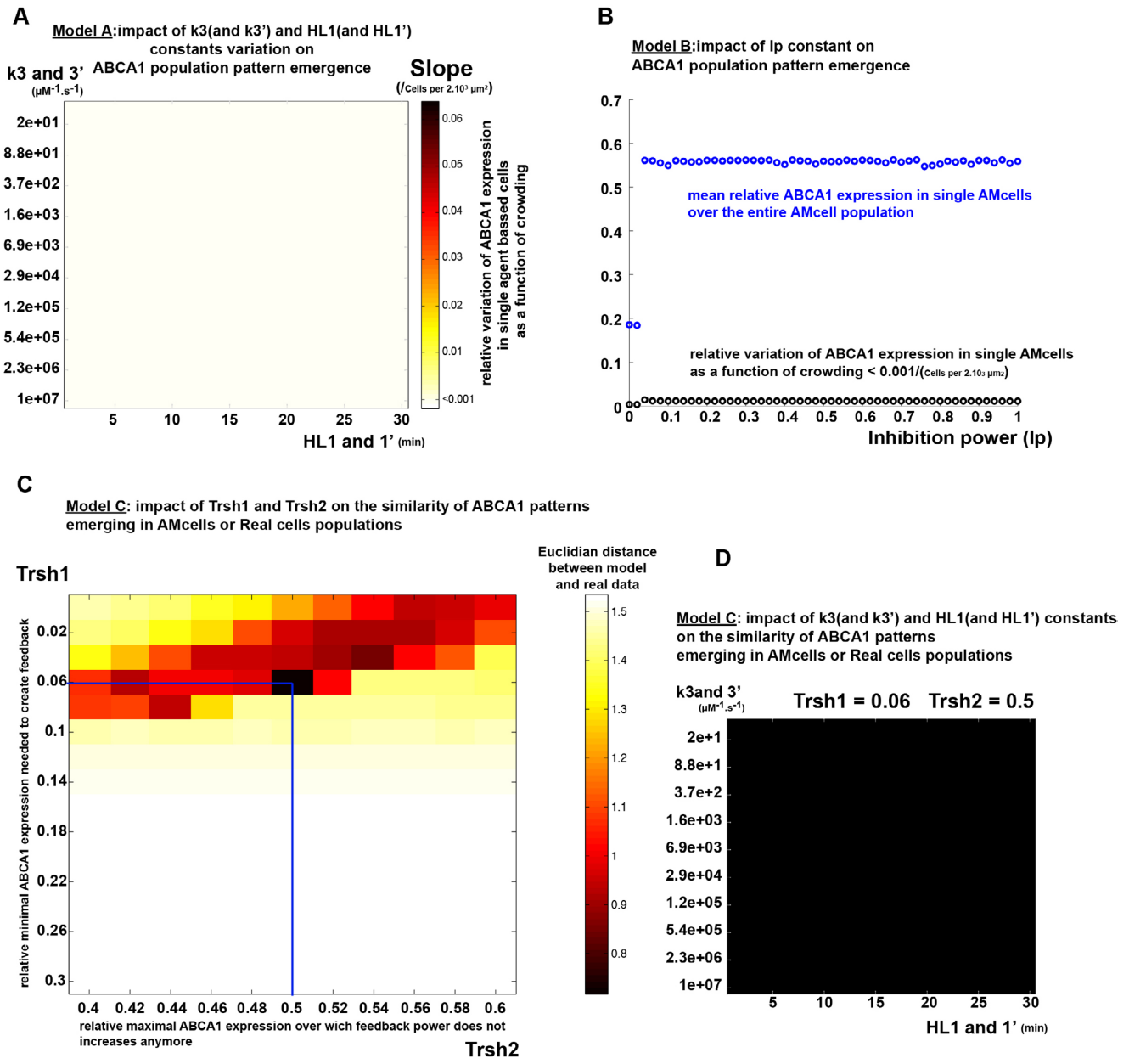
Extended Data Figure 3 | Agent-based modelled single cells show characteristics similar to tracked cells. **a**, Typical curve of the growth of the nucleus size of a single cell between two mitotic events (centre). Distribution of measured (number of tracks: 650) and agent-based modelled (number of tracks: 200) single-cell nucleus sizes (right histograms) and cell-cycle lengths (bottom histograms). Black, raw data, red, fitted Gaussian curve. Agent-based modelled cells and measured cells show similar distributions in cell-cycle length and nucleus size. **b**, Curve showing single-cell mean nuclear area against local cell crowding of measured (black, number of cells: $>10^4$) and agent-based modelled cells (red, number of cells: $>10^3$). **c**, Histograms of single-cell area distribution of measured (number of cells: $>10^4$) and agent-based modelled cells (number of cells: $>10^3$) showing that distribution of emerging cell areas of modelled cells are matching those of measured cells even for extreme values. **d**, Histograms of single-cell mean square displacement distribution of measured (number of tracks: 650) and agent-based modelled cells (number of tracks: 200). **e**, Timescales of information sensing and processing steps in the FAK–ABCA1 system. Absence of a capacitor does not allow gradual patterns to emerge (switch-like behaviour).



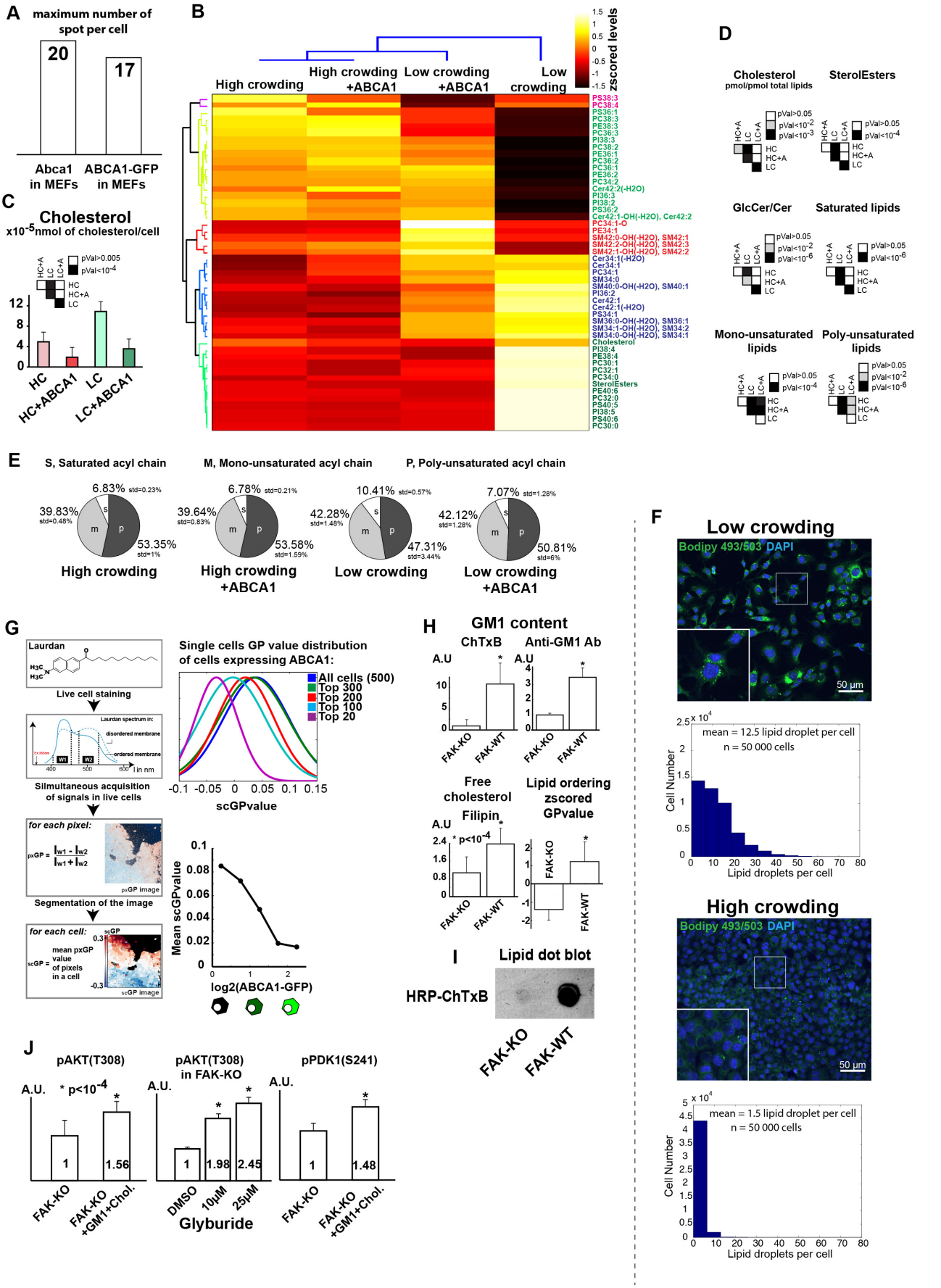
Extended Data Figure 4 | Alternative models do not lead to the emergence of gradual patterns in ABCA1 expression, and the full model recapitulates experimentally observed dynamics of reduction in ABCA1 expression in scratch assays. Conclusions are parameter-independent, for details see mathematical appendix in the Supplementary Information. **a**, A FAK activation model without autophosphorylation does not result in a pFAK pattern in an agent-based modelled cell population. **b**, A FAK–ABCA1 model based on free diffusion of signalling molecules without or with **c**, addition of a putative direct inhibitory effect of ABCA1 on its own suppression does not result in a patterning of ABCA1 expression. **d**, Introduction of a membrane relay for AKT activation without ABCA1 feedback on the membrane relay does not result in a patterning of ABCA1 expression. **e**, Simulated single-cell ABCA1 variability over local crowding is similar to the variability seen in our experiments (see Fig. 2d). **f**, Scratch assays, at which cells at high crowding suddenly become exposed to free space to spread and followed over time, show that reduction of ABCA1 levels in these cells has a half-maximum effect at ~50 min, and full effect at ~200 min. **g**, This is in agreement with simulations of scratch assays using our cell-intrinsic Agent-based model of the FAK–ABCA1 system. The process was iterated thousands of times with random starting levels of ABCA1 similar to the variability seen in the experimental scratch assay. 20 representative curves are shown. In the simulations, it takes ~150 min for the disappearance of half of ABCA1. **h**, Distributions of pixel GP values of FAK-KO cells stained with Laurdan at different time-points after treatment with glyburide. After just 20 min of drug treatment, the membranes of these cells become more ordered ($P < 10^{-100}$, *t*-test, pixel distributions at each time point are made from 2×10^3 cells).



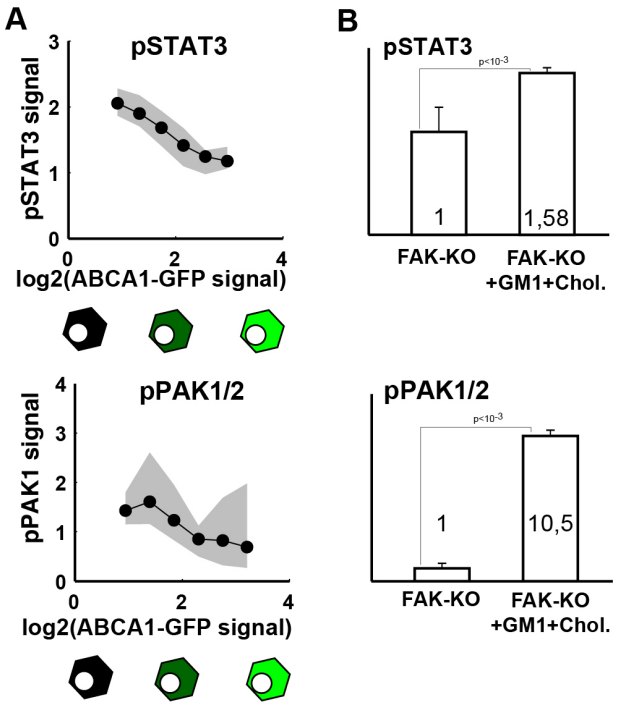
Extended Data Figure 5 | Sensitivity analysis of the FAK activation model. **a**, Heat map representing Euclidian distance between modelled and measured levels of pFAK in single cells as a function of local crowding when autophosphorylation constant k_1 and removal rate RR varies. Stars represent the values used for further modelling; any pair of k_1-RR values with the same low Euclidian distance will lead to the proper pFAK pattern. **b**–**d**, Same analysis for k_1 and the FAK-independent phosphorylation of FAK rate k_2 for a fixed RR value shows that FAK-independent phosphorylation of FAK has no effect on the formation of a pFAK pattern even if k_2 is bigger than k_1 by several orders of magnitude.



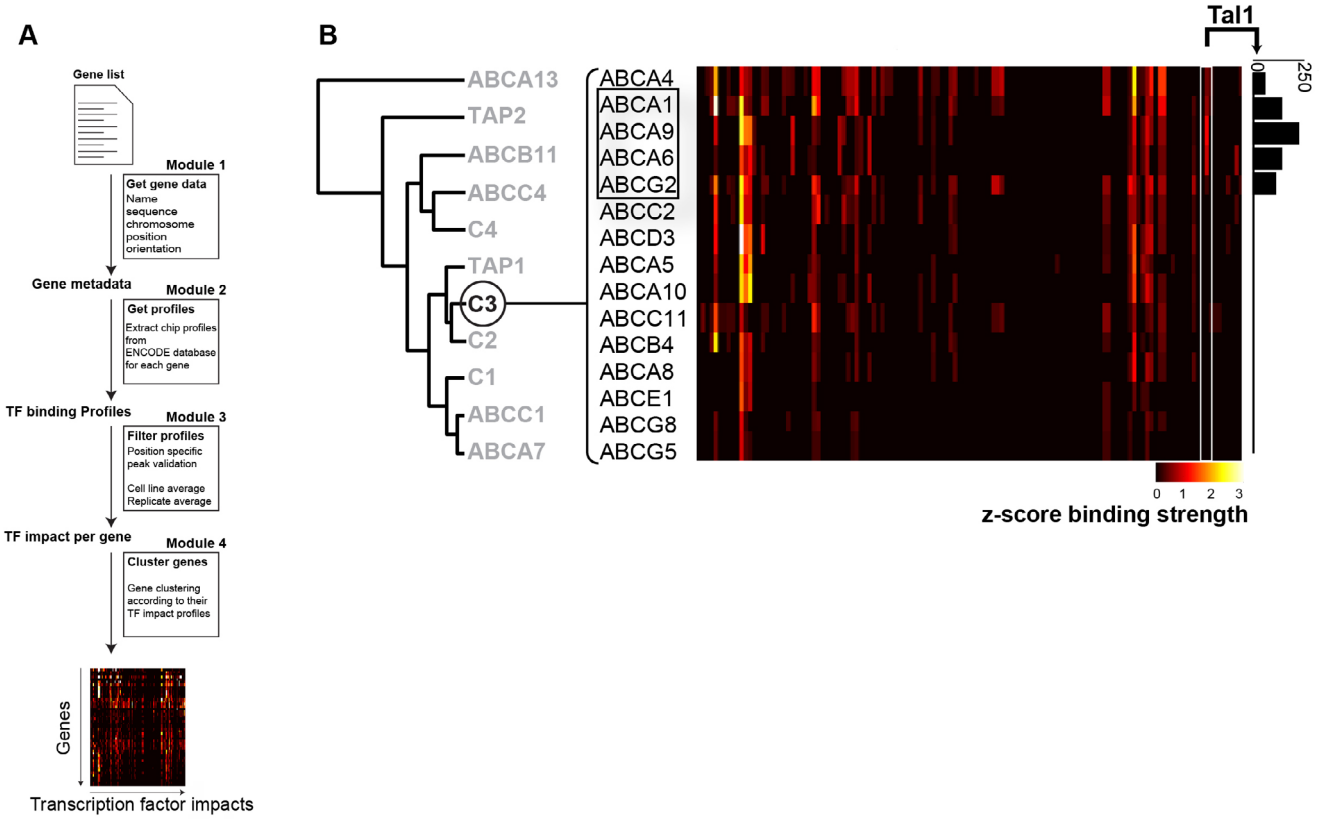
Extended Data Figure 6 | Sensitivity analysis of the FAK to ABCA1 expression models. **a**, Heat map representing the slope of ABCA1 expression against local cell crowding when k_3 and 3′ and HL1 and 1′ vary over an extreme range of values for model A. This demonstrates that such topology cannot lead to emergence of gradual expression patterns ABCA1 expression as a function of local cell crowding. **b**, Mean relative ABCA1 expression in agent-based modelled cells as a function of its inhibition power (Ip) in model B, where ABCA1 would be able to directly inhibit activation of AKT (or PI(3)K). This demonstrates that such direct feedback only leads to switch-like behaviour where ABCA1 is either expressed or not in all cells of the population, independent of local cell crowding. Inhibition power represents the ABCA1 competitive inhibitory power. **c**, Heat map representing Euclidian distance between modelled and measured levels of ABCA1 in single cells as a function of local crowding when Trsh1 and Trsh2 vary in model C. **d**, The capacity of model C to generate a gradual expression pattern (low Euclidian distance is black) does not depend on k_3 and 3′, and HL1 and 1′, demonstrating the central role of the membrane relay for gradual patterns to emerge.



Extended Data Figure 7 | The FAK–ABCA1 system adapts membrane lipid composition, ordering and signalling to local crowding. Related to Fig. 4. **a**, Histogram of transcript copy number (number of spots) per cell determined with bDNA single-molecule FISH against endogenous *Abca1* in cells at high crowding, or against *ABCA1–GFP* transcripts in cells at low crowding transfected with the pEGFP-N1-ABCA1 construct. This shows that plasmid-driven *ABCA1–GFP* expression in cells at low crowding does not exceed that of endogenous *Abca1* levels in cells at high crowding. **b**, Hierarchical clustering of lipid profiles of mouse embryonic fibroblasts grown at high crowding or low crowding conditions and transiently expressing ABCA1 from a plasmid (+ABCA1) or not. The clustergram shows the 48 lipid species that represent 80% of the total lipid amount. Colours correspond to pmol/pmol total lipid *z*-scored over the four conditions, colours of lipid names refer to their clusters. For complete lipid mass spectrometry data, see Supplementary Table 3. **c**, Histograms displaying the quantity of free cholesterol in nmol per cell ($n = 4$ biological replicates, each the mean of 4 technical replicates, s.d.). **d**, *P* values related to the bar graphs in Fig. 4c. **e**, Pie charts representing the percentage of saturated, monounsaturated and polyunsaturated lipids for the four different conditions. **f**, Fluorescence imaging using Bodipy 493/503 dye of lipid droplets in low crowding ($n = 5 \times 10^4$ single cells) or high crowding conditions ($n = 5 \times 10^4$ single cells). This confirms that cells at low crowding contain a larger amount of cholesteryl-esters, which are stored in lipid droplets. **g**, Diagram summarizing the method to measure membrane ordering of a formaldehyde fixed population of cells at the single-cell level (left flow chart). Distributions of single-cell GP values for groups of cells that are the top 20, 100, 200, 300 ABCA1–GFP expressing cells compared to all cells (top right distributions, $n = 500$ cells) and curve showing the relationship between single-cell ABCA1 expression and scGP value (bottom right curve, $n = 500$ cells). **h**, Image-based quantification of free cholesterol (filipin), GM1 content (cholera toxin B binding or anti-GM1 antibody) and lipid ordering (Laurdan, as in panel **d**) in single MEFs with (FAK-WT) or without FAK (FAK-KO). $n = 4$ experiments, each $>10^4$ cells. **P* values (*t*-test) $< 10^{-4}$. **i**, Because some GM1 may not be accessible in formaldehyde-fixed cells, we performed dot blot analysis of lipid extracts from FAK-KO and FAK-WT cells using HRP-conjugated cholera toxin B. This indicates that FAK-WT cells have higher levels of GM1 than FAK-KO cells. **j**, pAKT and pPDK1 immunostaining in cells without FAK (FAK-KO) exogenously loaded with GM1 and cholesterol (FAK-KO + GM1 + Chol.), treated with DMSO, or with 10 and 25 μM glyburide in DMSO ($n = 3$ experiments, each 10^4 cells, s.d., **P* values (*t*-test) $<10^{-4}$.



Extended Data Figure 8 | Phosphorylation of STAT3 and PAK1/2 are sensitive to ABCA1-mediated membrane perturbation. **a**, Curve showing the relationship between ABCA1–GFP expression and phosphorylated STAT3 (T705) and PAK1/2 (T423/T402) amounts in single cells. **b**, Quantification of immunostaining of phosphorylated STAT3 (T705) and PAK1/2 (T423/T402) amounts in FAK-KO cells after exogenous loading of the plasma membrane with cholesterol and GM1 (s.d., $n = 4$ experiments, each with 10^4 cells, *t*-test).



Extended Data Figure 9 | Hierarchical clustering of human ABC transporters according to 118 transcription factor binding profiles from the ENCODE database. **a**, Diagram of the algorithm used to generate ABC transporter clusters. **b**, Heat map of the cluster of ABC transporters containing ABCA1, A9, A6 and G1 that share Tal1 binding (see bar graph representation of Tal1 binding on the right). These 4 ABC transporters are the same 4 ABC transporters that were found higher expressed in cells lacking FAK (FAK-KO) (see Extended Data Fig. 1c).

doi:10.1038/nature14429

Cell Profiler modules developed for this study and computer code developed for the Agent-based modeling approach can be found at GitHub, an open-source repository of computer code and software: https://github.com/pelkmanslab/

1. Tracking

a. Movie acquisition using automated spinning disk confocal microscopy

Movies of A431 cells were acquired during 3 days in 96 well plates using Yokogawa CV7000 automated spinning disk microscope under environment control (37°C and 5% CO_2). The acquisition was done at a frequency of one image every 15 minutes. Cells stably expressing CAV1-GFP and live-stained with 500 nM of Hoechst were used for cell outline and nucleus detection purposes.

b. Applying Otsu thresholding, Laplacian of Gaussian filtering and propagation algorithm for automated single cell detection.

Detection and segmentation of objects and extraction of related measurements is done independently in each image set on our iBRAIN platform that runs CellProfiler [27] jobs in a parallelized fashion. Computation is done on the ETH Brutus cluster as described before [3,11,28]. Briefly, nuclei are detected and segmented based on Hoechst signal in IdentifyPrimaryObjects CellProfiler module. We first separate pixels in two classes, background and signal using

Otsu thresholding algorithm based on variance reduction [39] . We chose the Otsu method as it works consistently when number of objects within images of the dataset can vary significantly. Indeed from the beginning to the end of the movie, populations of cells grow and the number of objects within images increases significantly over time. Smooth detection of nuclei borders and declumping of objects is then performed using Laplacian of Gaussian filtering method [40].

Detected primary objects (nuclei) are then used as seeds for cell outline detection using IdentifySecondaryObjects CellProfiler module, based on CAV1-GFP signal. This module uses an improved watershed algorithm (propagation algorithm) [27,41], which combines information coming from the distance of close detected nuclei and gradient of surrounding signal intensity to detect local cell outlines. Nuclei and cell outlines are then used to extract object related measurements like nuclei and cell size, shape or intensities. Extensive documentation and code is available on CellProfiler web site (http://www.cellprofiler.org/). In this case, for tracking reasons, mitotic cells are not filtered out using support vector machine (SVM), like we did in previous studies [3,11,28], as mitotic events are essential for proper lineage detection.

c. Automated high content single-cell tracking

We tracked objects using a custom automated single cell tracker based on the TrackObjects CellProfiler module that works without manual correction steps. Indeed the relatively high frequency of acquisition of the movie (4 images/hour), and its good resolution allows us to track objects using a robust overlap method

with no minimization step involved, however we do take into account consistency in measurements over single cell traces in the latest filtering step of our algorithm. Due to efficient primary object detection, the tracking is very robust, and allowed us to track thousands of cells. We applied a stringent filter and selected ~700 cells tracked over at least one entire cell cycle (18-22 hours) and we use time resolved measurements, cell size, nuclear size, mean square displacement and length of cell cycle to test and validate the behavior of our agent-based models of cell proliferation and growth.

The code is available upon request.

2. Agent-based modeling of single cells to recapitulate population phenomena

Emergence of large-scale behavior from numerous small-scale interactions is a fascinating characteristic of complex systems. Understanding how complex biological functions emerge from the interplay of numerous simple mechanisms occurring within a cell is a major challenge in systems biology. Agent-based modeling is a tool of prime interest for answering this question where units or agents interacting in a predefined space responding to very simple rules, adopt a collective behavior that supports an emergent larger-scale organization that could not be predicted at the agent level [42].

The graphical and intuitive nature of agent-based models led to their frequent use in social and behavioral science, and economics [42], and to a certain

extent also in biology [43–46]. We designed here a two-level agent-based model that is at the first level made of agents simulating focal adhesion structures (FA) encapsulated within single cells, and at the second level made of agents simulating single cells, governed in part by the collective behavior emerging from the agents at the first level. Our agent-based modeled cells (simply called later ''cells'') are able to sense and react to local crowding, to spread, divide and migrate. While the behavior of single agents at the first level is largely random within the constraints given by the available space to move, we observe the emergence of self-organized cell populations that closely mimic populations of real cells (see supplementary Fig. 3). This agent-based model allows us to have virtual populations of cells behaving close to reality, on which we apply nested and responsive bottom-up models to decipher general rules about cell signaling control and adaptation of the cell physiology to external cues like local cell crowding.

a. Principles

Our agent-based model design is inspired by the literature on focal adhesions that can form, expand and mature if the cell has enough space to spread [6,47], but also by our Total Internal Reflection Fluorescence Microscopy (TIR-FM) movies of Hela cells stably expressing a GFP-Focal adhesion kinase construct (Supplementary movie 1), where one can see that focal adhesions are constantly probing the microenvironment by sampling the space accessible for the cell. Therefore, our agent-based model of a single cell is composed of two parts, the agents representing focal adhesions that are able to randomly move

and expand, and the nuclei agent that collects information from the focal adhesion agents and decides when to move and divide. The cell boundary is defined by linking the focal adhesion agents and the nuclei agent is always contained within the region delimited by the focal adhesion agents (Fig. 3a, supplementary movie 2).

Focal adhesion (FA) agents can randomly move in the adaptive and non-uniform potential intracellular landscape. This landscape encapsulates the FA agents and the nuclei agent into a single-cell agent and sets the potential maximal or minimal size it can reach according to real data extracted from our experiments, which can be regarded as an intrinsic limit for cytoskeleton expansion or contraction. The landscape of a cell can be altered by other cell agent boundaries close enough to create non-crossable barriers such that the movement of FA agents and cells respond to the local microenvironment (Supplementary movie 2).

Each FA agent possesses a proper energy that sets its available region within the intracellular landscape at the current time point. In this region, the FA agent can reach any position in a random fashion. This energy is evenly distributed from the cell's energy depot to all FA agents and is given back to the depot if no move is made.

b. The simulation algorithm

i. The simulation starts by evaluating which cells must divide, following a semi random behavior that triggers division after a cell cycle of 21 hours +/- 2 hours of standard deviation according to the cell cycle length variability extracted from real movies (Fig. 3a). If a cell divides, two new cells appear in the area occupied by the mother cell before division.

ii. The next step defines the new random position of each FA agent within each cell agent, taking into account positions of the other moving cells. The challenge is to avoid the iterative nature of the computation to induce a bias in their movement. First, the new position of each cell agent is evaluated according to the constraints imposed by the surrounding cells at time point t. Once all new positions are defined, some cells overlap with their neighbours. The algorithm treats each cell-cell overlap by resetting and rerunning the random placement of FA agents of the concerned cells until it finds a possible common solution. Once the positions of all FA agents are found, a new cell population topology appears and gives the image at t+1. This then serves as the starting image for the new time point simulation, etc.. For this study the time increment of the simulation is set at 15 minutes, similar to the frequency of acquisition of our real movies we use to benchmark the model (Fig. 3a).

Our modeling approach accurately mimics single-cell social behavior and self-organization in cell populations (see next paragraph) using only very simple rules: The energy distribution between FA agents is even and constant, and the

movement of the cell agent is defined by the sum of the nuclei agent-FA agent vectors. This is sufficient to generate emergent properties that mimic cell movements, directionality, cell size distributions, etc.. While more complex interplays between agents can be modeled, for instance based on the asymmetric distribution of energy between FA agents to favor or counteract symmetry breaking and leading edge formation, or based on more subtle rules of energy uptake or migration behavior, we here prefer the simplest set of rules that are sufficient to recapitulate real cell population phenomena.

c. Emergent behavior of our model and similarity with reality

We compared a set of parameters extracted from model-simulated cell populations to real data in order to assess if our agent-based model of single cells and cell population growth behaves close to reality. Cell cycle length and nuclear size distributions of the modeled cells are comparable to reality, as well as the nuclear size distribution as a function of cellular crowding (Supplementary Fig. 3, a and b). Also the cell area size distribution is similar between modeled and real cell populations, showing that proper cell area sizes can emerge from interactions between the modeled FA agents, even for very large cell area sizes (Supplementary Fig. 3c). Finally, the mean square displacement (MSD) distribution of the modeled cells accurately reproduces experimentally measured MSD distributions (Supplementary Fig. 3d), even for high displacement values, showing that in the simulations single cells with directional migration emerge. This occurs when FAs on one side of a cell face constraints due to the presence of neighbouring cells while on the other side

they do not face such constraints. This can lead to symmetry breaking, with FAs on the side where spreading is unobstructed building up more adhesion potential, leading to more activated FAK on that side. Since cell migration is determined by the net vector of the distances of FAs to the nucleus of a single cell, these cells will start to migrate into the direction of unobstructed spreading.

3.Mathematical appendix

a. General remarks

The inherent robustness [48] of biology implies that processes show comparable dynamics despite a certain amount of fluctuation in total protein amounts [49].We therefore assume that degradation and synthesis rates of proteins are equal and that protein amounts of our different players are comparable and around 1μM [50] in each modeled single cell. Thus, values describing phosphorylated or unphosphorylated forms of the proteins can be seen as relative proportions, improving the readability and simplicity of our models.

For all models:

[PROTEIN] = 1-[pPROTEIN]

The conclusions that we draw from the dynamic behavior of our models are similar even when the signaling strength oscillates widely over time scales from milliseconds to minutes due to intrinsic noise [51]. Thus, for clarity and readability of our conclusions such extra levels of complexity are not modeled.

b. Mathematical model for simulation of FAK activation patterns

We have nested a mathematical model of focal adhesion kinase activation (FAK) within each FA agent of each cell agent, where activation of FAK is promoted by an increase in the FA agent adhesion potential. The goal is to favor the emergence of an active FAK pattern similar to what we observe within populations of real single cells (Fig. 3b, supplementary movie 3) in reaction to cell agent growth and proliferation and self-organization into cell populations.

Our FAK activation model **(a)** is based on first and second order kinetic reaction simulation and in three parts summarizes the knowledge accumulated on the FAK activation cycle with no *a priori* constraints on each related constant. Values for these constants are extracted after fitting our modeled population patterns with real active FAK patterns (Supplementary table 4)(Supplementary Fig. 5).

$$\frac{d[\mathrm{pFAK}]}{dt} = k_1 \frac{[\mathrm{pFAK}][\mathrm{FAK}]}{e^{(-FA)}} + k_2[\mathrm{FAK}] - RR\frac{[\mathrm{pFAK}]}{e^{(FA)}} \qquad (a)$$

The first part of the dynamic model simulates the autophosphorylation capacity of pFAK that occurs in focal adhesions (FAs) and therefore is promoted by FA growth [52,53] with second order kinetic constant k_1. The second part of the equation represents FAK independent FAK activation [54] with first order kinetics constant k_2, and the third part of the model simulates the dephosphorylation of active FAK by phosphatases promoted by the reduction of FAs and driven by first order kinetics constant RR [6,55]. The constants required for simulating real active FAK patterns (Fig. 3b, supplementary Fig 5, supplementary table 4) tell us that autophosphorylation requires strong local enrichment in FAs with a slow constant k_1 compared to usual reported kinase phosphorylation constants that are mainly diffusion limited [56,57]. Here k_1 is in the order of magnitude of an active nuclear import/export constant [57] which can be explained by the strong entrapment of FAK in the mesh of proteins building FAs, keeping active FAK away from phosphatases but limiting its diffusion, in accordance with published work [52,58]. Thus in our model, FAK activation rate is limited by its recruitment in the FA.

Moreover, in the model, the driving force for FAK activation is its autophosphorylation capacity (supplementary Fig. 4a), which cannot be accounted for by independent FAK activation even if the related constant k_2 is three orders of magnitude higher than k_1 (supplementary Fig. 5, b-d). This observation fits the current literature, where the autophosphorylation of FAK at Tyr-397 is mandatory for further phosphorylation to occur [59], suggesting that autophosphorylation is what drives FAK signaling activation.

It would be intuitive to include a feedback between activated FAK and the behavior of FAs modeled with the ABM, such as their disassembly and abundance. This is technically very challenging as it massively increases the amount of computing time needed for simulations. For the purpose of this study, such a more advanced model would not change the outcome of the patterning, as it merely strengthens the properties of FAK auto-activation by including a second positive feedback acting via FAs themselves. It would however be an interesting further direction to take when studying details of cell migration within cell populations.

c. Mathematical model for ABCA1 production

Our experiments demonstrate that FAK and ABCA1 expression are linked through PI3K-AKT signaling, resulting in ABCA1 expression patterns across a population of cells (Fig. 2d). Thus, our goal was to recreate such a pattern of ABCA1 expression in agent-based modeled cell populations. Using a step-by-step approach (see below), we realized a cell-intrinsic mechanism like the FAK-ABCA1 system can only generate a gradual pattern of ABCA1 expression as a function of local cell crowding when signaling information flow carried by protein phosphorylation events, which is intrinsically very fast [56,57,60], is integrated over a longer time-scale, minutes to hours, with the help of a 'relaying' structure. The cellular counterpart of this relaying structure is the plasma membrane, which acts as a storage of information generated by PI3K (namely PIP_3), and transforms it into a longer time-scale by imposing that PIP_3

production and accumulation must occur during a considerable time before sufficient PIP_3 is generated and activation of AKT by PDK1 occurs (see also Fig. 3d).

i. Second order free diffusion kinetic system fails to explain pattern formation (Supplementary Fig. 4b)

The first model we developed was made of a simple cascade of protein phosphorylation events aiming at representing in a simple way the information flow from FAK to ABCA1 production, using second order kinetics. Phosphorylation of PI3K and AKT are driven by second order kinetic constants k_3 and k_3' and half-lives HL_1 and HL_1' that are defined according to kinetic studies carried out previously, which show that intracellular kinetic constants are diffusion limited [56,57], and reported half-lives of a few minutes for phospho-proteins[57,61]. The production rate of ABCA1 is then linearly and inversely linked to phospho-AKT levels with a maximal production rate PR [62] multiplied by maximal mRNA count and an half-life HL_2 [63–65] (table S5).

Model A: **(*a*)** is evaluated for each FA agent in each single cell; **(*b, c, d*)** are evaluated for each cell agent.

$$\frac{d[\mathrm{pFAK}]}{dt} = k_1 \frac{[\mathrm{pFAK}][\mathrm{FAK}]}{e^{(-FA)}} + k_2[\mathrm{FAK}] - RR\frac{[\mathrm{pFAK}]}{e^{(FA)}} \qquad (a)$$

$$\frac{d[\mathrm{pPI3K}]}{dt} = k_3 \sum [\mathrm{pFAK}][\mathrm{PI3K}] - [\mathrm{pPI3K}]0.5^{\left(\frac{\Delta t}{HL_1}\right)} \qquad (b)$$

$$\frac{d[\text{pAKT}]}{dt} = k_3,[\text{pPI3K}][\text{AKT}] - [\text{pAKT}]0.5^{\left(\frac{\Delta t}{HL_1,}\right)} \quad \text{(c)}$$

$$\frac{d[\text{ABCA1}]}{dt} = PR\ [\text{1-pAKT}] - [\text{ABCA1}]0.5^{\left(\frac{\Delta t}{HL_2}\right)} \quad \text{(d)}$$

While in reality a cascade of protein phosphorylation is more complex than in the model, our conclusions are insensitive to how detailed and complex the cascade is, as well as to variations of kinetics constants over seven orders of magnitude around the commonly reported values, and a wide range of half lives around commonly reported values (Supplementary Fig. 5a). Given the failure of this model to reproduce ABCA1 patterns, this indicates that a different topology, and not necessarily more complexity in the current topology, is necessary to reproduce ABCA1 patterning.

ii. A homogeneous second order kinetic system with competitive inhibition feedback fails to explain pattern formation (Supplementary Fig. 4c)

From the results above, it seemed plausible that some self-amplifying effect may be necessary in order to achieve sufficient expression of ABCA1 in cells experiencing high local crowding. Since some reports have shown that ABCA1 binds to and inhibits signaling proteins [66], we changed the topology of our FAK to ABCA1 ***Model A*** by introducing in equation **(c)** a potential capacity of ABCA1 to directly bind AKT (or PI3K), acting as a competitive inhibitor for phospho-AKT

(or phospho-PI3K) production. In essence, this constitutes a double-negative (i.e. positive) feedback loop of ABCA1 on itself. We modeled this (***Model B***) by introducing an inhibition power constant I_p by which ABCA1 inhibits AKT (or PI3K) to enter into a productive activation complex. Variations in this constant represent possible different inhibition strengths of ABCA1 as well as possible differences in relative protein/protein ratios.

Model B is similar to Model A with equation (c) replaced by equation (e):

$$\frac{d[\text{pFAK}]}{dt} = k_1 \frac{[\text{pFAK}][\text{FAK}]}{e^{(-FA)}} + k_2[\text{FAK}] - RR\frac{[\text{pFAK}]}{e^{(FA)}} \qquad (a)$$

$$\frac{d[\text{pPI3K}]}{dt} = k_3 \sum [\text{pFAK}][\text{PI3K}] - [\text{pPI3K}]0.5^{\left(\frac{\Delta t}{HL_1}\right)} \qquad (b)$$

$$\frac{d[\text{pAKT}]}{dt} = k_{3'}[\text{pPI3K}]([\text{AKT}] - [\text{ABCA1}]I_p) - [\text{pAKT}]0.5^{\left(\frac{\Delta t}{HL_{1'}}\right)} \qquad (e)$$

$$\frac{d[\text{ABCA1}]}{dt} = PR\ [1\text{-pAKT}] - [\text{ABCA1}]0.5^{\left(\frac{\Delta t}{HL_2}\right)} \qquad (d)$$

Running simulations over a wide range of I_p values revealed that a double-negative feedback modeled in this way does **not** result in gradual patterns of ABCA1 expression as a function of local cell crowding. Rather, depending on the

inhibition power I_p of ABCA1, we observed either no ABCA1 expression (if I_p is too small, Model B is equal to Model A)(Supplementary Fig. 4b) or an equally high ABCA1 expression in all single cells across the population despite variations in local cell crowding (Supplementary Fig. 4c, supplementary table 6). This ''all or nothing behavior'' is explained by the fact that, by competing directly with a phosphorylation reaction that is governed by free diffusion, ABCA1 acts within the same time-scale (milliseconds), and therefore induces a fast switching response compared to the process of cell spreading, proliferation and population growth leading to changes in local cell crowding, which occurs at the minute and hour time-scale.

iii. A membrane-based signaling relay and feedback integrates time-scales and generates gradual patterns of ABCA1 expression in cell populations similar to reality (Fig. 3c, supplementary movie 3).

AKT activation happens on the membrane[67]. This requires the membrane to be enriched in PIP_3 in order to recruit AKT and its activator PDK1[68], which both bind to PIP_3 via a PH domain. In addition, the probability of AKT and PDK1 to meet on the membrane is enhanced when diffusion of PIP_3 in the membrane upon its production is low, which can be achieved by increasing the amounts of cholesterol and sphingolipids in the membrane[69], as this impacts membrane lipid ordering [70–73]. We thus added new steps in our model that represent a membrane 'relay', aiming at simulating a platform that stores PIP_3 upon production based on experimentally determined kinetics of PIP_3 production by

PI3K[74], and allowing ABCA1 to perturb general membrane ordering and thus diffusion of PIP_3 [16,19,75], thereby reducing the capacity of the membrane 'relay' to activate AKT **(*i*)**. PIP_3 production responds to first order kinetic constant k_4 and has a half-life HL_3 **(*f*)** defined in a previous study [74]**.** k_4 is expressed in minutes and HL_3 in hours, both much slower than protein phosphorylation constants (milliseconds). Levels of ABCA1 impact the capacity of AKT to get activated by PIP_3 with two specific thresholds, Trsh1 and Trsh2, which respectively represent the amount of ABCA1 required to trigger an effect, and the amount of ABCA1 with maximum effect, beyond which no stronger inhibitory effect is achieved **(*g*)**. This ''two-threshold'' modeling of the effect of ABCA1 on the capacity of the membrane to allow AKT activation follows from the Stokes-Einstein equation extended to lipid bilayers by Saffman and Delbuck[17]. The lower threshold reflects a minimal amount of perturbation (energy) conferred by ABCA1 to the membrane to increase diffusion, while the higher threshold reflects the maximal amount of perturbation possible, beyond which lipid diffusion cannot be further increased. Although this can be explicitly modeled, the two-threshold approach simplifies this part without changing the outcome. Since the model assumes that the time-scale of this feedback is mainly determined by the production rate of ABCA1 (which in de model is 40 proteins per hour, thus taking several hours for an effect to establish), we experimentally measured the rate of change in membrane ordering in FAK-ko cells (which express high levels of ABCA1) treated with Glyburide (ABCA1 inhibitor). This shows that drug-mediated inhibition of ABCA1 triggers a fast reordering of the membrane (20 minutes) indicating that once in the membrane ABCA1 acts fast. Therefore, the capacity of the system to perturb the membrane at a global scale is likely limited by the

production of ABCA1 rather than by the capacity of ABCA1 to work once in the membrane (Supplementary Fig. 4h).

Finally, we implemented a controller saturation limit, which simulates a capacity of the system to overcome ABCA1 inhibition in case of a strong reactivation of FAK signaling. If PIP_3 amounts overcome the threshold Trsh3, AKT can be reactivated even if ABCA1 inhibition acts at 100% **(*h*)**.

Model C: **(*a*)** is evaluated for each FA agent in each single cell; **(*b, f, g, h, i, d*)** are evaluated for each single cell agent.

$$\frac{d[\mathrm{pFAK}]}{dt} = k_1 \frac{[\mathrm{pFAK}][\mathrm{FAK}]}{e^{(-FA)}} + k_2[\mathrm{FAK}] - RR\frac{[\mathrm{pFAK}]}{e^{(FA)}} \qquad (a)$$

$$\frac{d[\mathrm{pPI3K}]}{dt} = k_3 \sum [\mathrm{pFAK}][\mathrm{PI3K}] - [\mathrm{pPI3K}]0.5^{\left(\frac{\Delta t}{HL_1}\right)} \qquad (b)$$

$$\frac{d[\mathrm{PIP3}]}{dt} = k_4[\mathrm{pPI3K}] - [\mathrm{PIP3}]0.5^{\left(\frac{\Delta t}{HL_3}\right)} \qquad (\mathrm{f})$$

$$\mathrm{FBUnit} = [\mathrm{PIP3}] - \frac{[ABCA1] - Trsh_1}{Trsh_2 - Trsh_1} \qquad (g)$$

$$\mathrm{OSUnit} = \frac{[\mathrm{PIP3}] - Trsh_3}{1 - Trsh_3} \qquad (h)$$

$$[\text{pAKT}] = \text{FBUnit} + \text{OSUnit} \qquad (i)$$

$$\frac{d[\text{ABCA1}]}{dt} = PR\ [\text{1-pAKT}] - [\text{ABCA1}]0.5^{\left(\frac{\Delta t}{HL_2}\right)} \qquad (d)$$

By modeling the signaling and the double-negative feedback of ABCA1 on its own expression in this way, the model allows simulating gradual patterns of ABCA1 expression in a cell population as a function of local cell crowding, similar to real patterns (Fig. 3c). Importantly, only the constants related to the mathematical definition of the membrane relay (data not shown) and the feedback through an alteration of the membrane (Supplementary table 5 and 6), are determining the ability to simulate a realistic ABCA1 expression pattern (Supplementary Fig. 4d; Supplementary Fig. 6, c and d). This shows that a key aspect of the patterning system is the role of the membrane and membrane lipid ordering to allow upstream signaling information and feedback control to be processed at the right time scale at which changes in the extrinsic stimulus occur, namely, in this case, local cell crowding.

The model was designed to unravel the global topology of our molecular system, and not the fine details. There is undoubtedly more complexity to take into account. However, its simplicity reveals the importance of time-scale control

of information flow, something that can be achieved by membrane-based signaling. The model also shows that a cell-intrinsic system can generate gradual patterns at the cell population level without the need for gradients of secreted signaling molecules, if the time-scale of intracellular information processing is adapted to the time-scale of cell population phenomena such as the emergence of differences in local cell crowding. One can imagine that more complex cell population patterns can emerge from other more elaborate topologies that remain to be unraveled.

4.Supplementary discussion

An unbiased clustering of human ABC transporters based on their TF-binding profiles from ENCODE[76] revealed that *ABCA1*, *-6*, *-9*, and *ABCG2* form a sub-cluster based on their shared binding to Tal1 (Supplementary Fig. 9). These are the same 4 ABC transporters suppressed by FAK in mouse embryonic fibroblasts experiencing low crowding (see supplementary Fig. 1c), suggesting that Tal1 is a key component in adaptation of ABC transporter expression across mammalian cells.

Membrane lipid composition impacts many cellular properties including the permeability, rigidity, and tension of membranes, membrane protein structure and function, the cytoskeleton, and membrane trafficking[77–80]. This suggests a fundamental role for the FAK-ABCA1 system in controlling cell behaviour within a social context, including the adaptation of patterning systems that rely on secreted molecules and cell surface receptors. Since ABC transporters are ubiquitous across organisms, they may be generally used to create patterns of cellular behaviour within a cell collective, conserved from colony formation in unicellular organisms[81,82] to collective cell migration and the patterning of intracellular tension and tension sensing

in epithelia[83,84], as well as stem cell differentiation in multicellular organisms[24]. In addition, since many ABC transporters can also transport xenobiotic drugs, population context-determined pattern formation in their expression may underlie the emergence of heterogeneous drug sensitivities amongst single cells[85].

References and Notes

39. Sezgin, M. & Sankur, B. Survey over image thresholding techniques and quantitative performance evaluation. *J. Electron. Imaging* **13,** 146–168 (2004).

40. Jain, R., Kasturi, R. & Schunck, B. G. Machine vision. 32018 (1995).

41. Jones, T. R., Carpenter, A. & Golland, P. Voronoi-Based Segmentation of Cells on Image Manifolds.

42. Billari, F. C. *Agent-based computational modelling: applications in demography, social, economic and environmental sciences*. (Taylor & Francis, 2006).

43. Vedel, S., Tay, S., Johnston, D. M., Bruus, H. & Quake, S. R. Migration of cells in a social context. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 129–34 (2013).

44. Solovyev, A., Mi, Q., Tzen, Y.-T., Brienza, D. & Vodovotz, Y. Hybrid equation/agent-based model of ischemia-induced hyperemia and pressure ulcer formation predicts greater propensity to ulcerate in subjects with spinal cord injury. *PLoS Comput. Biol.* **9,** e1003070 (2013).

45. Mukhopadhyay, R. *et al.* Promotion of variant human mammary epithelial cell outgrowth by ionizing radiation: an agent-based model supported by in vitro studies. *Breast Cancer Res.* **12,** R11 (2010).

46. An, G., Mi, Q., Dutta-Moscato, J. & Vodovotz, Y. Agent-based models in translational systems biology. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **1,** 159–71 (2009).

47. Geiger, B., Spatz, J. P. & Bershadsky, A. D. Environmental sensing through focal adhesions. *Nat. Rev. Mol. Cell Biol.* **10,** 21–33 (2009).

48. Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J. & Doyle, J. Robustness of cellular functions. *Cell* **118,** 675–85 (2004).

49. Maheshri, N. & O'Shea, E. K. Living with noisy genes: how cells function reliably with inherent variability in gene expression. *Annu. Rev. Biophys. Biomol. Struct.* **36,** 413–34 (2007).

50. Legewie, S., Herzel, H., Westerhoff, H. V & Blüthgen, N. Recurrent design patterns in the feedback regulation of the mammalian signalling network. *Mol. Syst. Biol.* **4,** 190 (2008).

51. Ladbury, J. E. & Arold, S. T. Noise in cellular signaling pathways: causes and effects. *Trends Biochem. Sci.* **37,** 173–8 (2012).

52. Brami-Cherrier, K. *et al.* FAK dimerization controls its kinase-dependent functions at focal adhesions. *EMBO J.* 1–15 (2014). doi:10.1002/embj.201386399

53. Schaller, M. D. *et al.* Autophosphorylation of the focal adhesion kinase, pp125FAK, directs SH2-dependent binding of pp60src. *Mol. Cell. Biol.* **14,** 1680–8 (1994).

54. Xing, Z. *et al.* Direct interaction of v-Src with the focal adhesion kinase mediated by the Src SH2 domain. *Mol. Biol. Cell* **5,** 413–21 (1994).

55. Kim, B., van Golen, C. M. & Feldman, E. L. Degradation and dephosphorylation of focal adhesion kinase during okadaic acid-induced apoptosis in human neuroblastoma cells. *Neoplasia* **5,** 405–16 (2003).

56. Fersht, A. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding.* (Macmillan, 1999).

57. Aoki, K., Yamada, M., Kunida, K., Yasuda, S. & Matsuda, M. Processive phosphorylation of ERK MAP kinase in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 12675–80 (2011).

58. Kanchanawong, P. *et al.* Nanoscale architecture of integrin-based cell adhesions. *Nature* **468,** 580–4 (2010).

59. Parsons, J. T. Focal adhesion kinase: the first ten years. *J. Cell Sci.* **116,** 1409–1416 (2003).

60. Aoki, K., Takahashi, K., Kaizu, K. & Matsuda, M. A quantitative model of ERK MAP kinase phosphorylation in crowded media. *Sci. Rep.* **3,** 1541 (2013).

61. Baker, A. F. *et al.* Stability of Phosphoprotein as a Biological Marker of Tumor Signaling umor Signaling. 4338–4340 (2005).

62. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–42 (2011).

63. Arakawa, R. & Yokoyama, S. Helical apolipoproteins stabilize ATP-binding cassette transporter A1 by protecting it from thiol protease-mediated degradation. *J. Biol. Chem.* **277,** 22426–9 (2002).

64. Wang, Y. & Oram, J. F. Unsaturated fatty acids inhibit cholesterol efflux from macrophages by increasing degradation of ATP-binding cassette transporter A1. *J. Biol. Chem.* **277,** 5692–7 (2002).

65. Wang, N. & Chen, W. A PEST sequence in ABCA1 regulates degradation by calpain protease and stabilization of ABCA1 by apoA-I. *J. Clin. …* **111,** 99–107 (2003).

66. Okuhira, K. *et al.* Binding of PDZ-RhoGEF to ATP-binding cassette transporter A1 (ABCA1) induces cholesterol efflux through RhoA activation and prevention of transporter degradation. *J. Biol. Chem.* **285,** 16369–77 (2010).

67. Manning, B. D. & Cantley, L. C. AKT/PKB signaling: navigating downstream. *Cell* **129,** 1261–74 (2007).

68. Alessi, D. R. *et al.* Characterization of a 3-phosphoinositide-dependent protein kinase which phosphorylates and activates protein kinase Balpha. *Curr. Biol.* **7,** 261–9 (1997).

69. Pike, L. J. & Miller, J. M. Cholesterol Depletion Delocalizes Phosphatidylinositol Bisphosphate and Inhibits Hormone-stimulated Phosphatidylinositol Turnover. *J. Biol. Chem.* **273,** 22298–22304 (1998).

70. Filippov, A., Orädd, G. & Lindblom, G. The effect of cholesterol on the lateral diffusion of phospholipids in oriented bilayers. *Biophys. J.* **84,** 3079–86 (2003).

71. Orlach, J. O. K., Chwille, P. E. S., Ebb, W. A. T. T. W. W. & Eigenson, G. E. W. F. Characterization of lipid bilayer phases by confocal microscopy. **96,** 8461–8466 (1999).

72. Rameh, L. E. The Role of Phosphoinositide 3-Kinase Lipid Products in Cell Function. *J. Biol. Chem.* **274,** 8347–8350 (1999).

73. Golebiewska, U., Nyako, M., Woturski, W., Zaitseva, I. & McLaughlin, S. Diffusion coefficient of fluorescent phosphatidylinositol 4,5-bisphosphate in the plasma membrane of cells. *Mol. Biol. Cell* **19,** 1663–9 (2008).

74. Auger, K. R., Serunian, L. a, Soltoff, S. P., Libby, P. & Cantley, L. C. PDGF-dependent tyrosine phosphorylation stimulates production of novel polyphosphoinositides in intact cells. *Cell* **57,** 167–75 (1989).

75. Denis, M., Landry, Y. D. & Zha, X. ATP-binding cassette A1-mediated lipidation of apolipoprotein A-I occurs at the plasma membrane and not in the endocytic compartments. *J. Biol. Chem.* **283,** 16178–86 (2008).

76. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489,** 91–100 (2012).

77. Laganowsky, A. *et al.* Membrane proteins bind lipids selectively to modulate their structure and function. *Nature* **510,** 172–175 (2014).

78. Lingwood, D. & Simons, K. Lipid rafts as a membrane-organizing principle. *Science* **327,** 46–50 (2010).

79. Rawicz, W., Olbrich, K. C., McIntosh, T., Needham, D. & Evans, E. Effect of chain length and unsaturation on elasticity of lipid bilayers. *Biophys. J.* **79,** 328–39 (2000).

80. Loose, M., Fischer-Friedrich, E., Herold, C., Kruse, K. & Schwille, P. Min protein patterns emerge from rapid rebinding and membrane interaction of MinE. *Nat. Struct. Mol. Biol.* **18,** 577–83 (2011).

81. Zhu, X. *et al.* A putative ABC transporter is involved in negative regulation of biofilm formation by Listeria monocytogenes. *Appl. Environ. Microbiol.* **74,** 7675–83 (2008).

82. Hinsa, S. M., Espinosa-Urgel, M., Ramos, J. L. & O'Toole, G. a. Transition from reversible to irreversible attachment during biofilm formation by Pseudomonas fluorescens WCS365 requires an ABC transporter and a large secreted protein. *Mol. Microbiol.* **49,** 905–918 (2003).

83. Dupont, S. *et al.* Role of YAP/TAZ in mechanotransduction. *Nature* **474,** 179–83 (2011).

84. Sinha, B. *et al.* Cells respond to mechanical stress by rapid disassembly of caveolae. *Cell* **144,** 402–13 (2011).

85. Singh, D. K. *et al.* Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities. *Mol. Syst. Biol.* **6,** 369 (2010).

Supplementary table 2: summary of the reasons for selecting candidate transcription factors

	Enrichment analysis for all FAK suppressed genes		**ABCA1 specific literature**	
	GeneGo	Pscan	ABCA1 agonism	Binding site
FoxA2				ChIP
FoxI1		x		
FoxO1				
FoxO3	x	x		ChIP
Max			x	x
Myc	x		x	x
Lxr-beta			x	ChIP
Lxr-alpha			x	ChIP
Nr3c1		x	x	x
Pparg			x	x
RelA	x	x	x	
Stat1	x	x		
Stat2				
Stat3	x	x		Luciferase
Stat4				Luciferase
Stat5a	x			
Stat5b	x			
Stat6	x			
Tal1		x		ChIP

Supplementary table 4: pFAK model constants

Constant	**Equation #**	**Related reaction**	**Value**	**Unit**	**Impact on pFAK pattern formation?**	**comment**
k_1	*(a)*	FAK ⇌ pFAK	8.10^{-3}	/M/sec	Yes	Supp. Fig. 5a
k_2	*(a)*	FAK ⇌ pFAK	-inf to 2.10^{1}	/M/sec	No	Supp. Fig. 5b-d
RR	*(a)*	FAK ⇌ pFAK	5.10^{-4}	/sec	Yes	Supp. Fig.

	5a

Supplementary table 5: ABCA1 expression model constants

Constant	Equation #	Related reaction	Value tested (Commonly reported value)	Unit	Impact on ABCA1 pattern formation?	Comment
k_3	*(b)*	PI3K ⇌ pPI3K	$10^1$to $10^7(10^4)$	/μM/sec	No	Supp. Fig. 6a,d [56,57]
$k_{3'}$	*(c)(e)*	AKT ⇌ pAKT	$10^1$to $10^7(10^4)$	/μM/sec	No	Supp. Fig. 6a,d [56,57]
HL_1	*(b)*	PI3K ⇌ pPI3K	1 to 30(5)	Min	No	Supp. Fig. 6a,d [57,61]
$HL_{1'}$	*(c)(e)*	AKT ⇌ pAKT	1 to 30(5)	Min	No	Supp. Fig. 6a,d [57,61]
HL_2	*(d)*	ABCA1 prod.	2	Hrs	No	63–65
PR	*(d)*	ABCA1 prod.	40*	Protein/Hrs	No	62
k_4	*(f)*	PIP3 ⇌ pPIP3	0.065	/Min	Yes	74
HL_3	*(f)*	PIP3 ⇌ pPIP3	1	/Hrs	Yes	74

* This value is calculated by the multiplication of the maximal ABCA1 mRNA count found in the group of most crowded cells multiplied by the rate of protein synthesis generally reported for large proteins in [62]. This value has no consequence on the behavior of the model (data not shown)

Supplementary table 6: Membrane and feedback related constant

Constant	Equation #	Related reaction	Value	Unit	Impact on ABCA1 pattern formation?	comment
$Trsh_1$	*(g)*	Membrane feedback	6	%	Yes	Supp. Fig. 6c
$Trsh_2$	*(g)*	Membrane feedback	50	%	Yes	Supp. Fig. 6c
$Trsh_3$	*(h)*	Membrane feedback	99	%	Yes	Supp. Fig. 6c
I_p	*(e)*	Direct feedback	0-100	%	No	Supp. Fig. 6b

5. Image-based transcriptomics in thousands of single human cells at single-molecule resolution.

By

Nico Battich*, Thomas Stoeger* & Lucas Pelkmans.

Published in *Nature Methods,* 21 March 2013.

doi:10.1038/nmeth.2657

*Contributed equally.

All experiments described in this chapter were designed, conducted, and analyzed in equal contribution by Nico Battich and Thomas Stoeger. The text of this chapter was written in equal contribution by Nico Battich and Thomas Stoeger.

Image-based transcriptomics in thousands of single human cells at single-molecule resolution

Nico Battich[1–3], Thomas Stoeger[1–3] & Lucas Pelkmans[1]

Fluorescence *in situ* hybridization (FISH) is widely used to obtain information about transcript copy number and subcellular localization in single cells. However, current approaches do not readily scale to the analysis of whole transcriptomes. Here we show that branched DNA technology combined with automated liquid handling, high-content imaging and quantitative image analysis allows highly reproducible quantification of transcript abundance in thousands of single cells at single-molecule resolution. In addition, it allows extraction of a multivariate feature set quantifying subcellular patterning and spatial properties of transcripts and their cell-to-cell variability. This has multiple implications for the functional interpretation of cell-to-cell variability in gene expression and enables the unbiased identification of functionally relevant *in situ* signatures of the transcriptome without the need for perturbations. Because this method can be incorporated in a wide variety of high-throughput image-based approaches, we expect it to be broadly applicable.

Large-scale transcriptomics with microarrays or RNA-seq is usually applied on a population of RNA molecules pooled from a large number of cells[1–4]. Although sequencing of single-cell transcriptomes has been performed[5–8], current approaches work reliably only for abundant RNAs[9], are feasible for only a small number of single cells and do not reveal the subcellular localization of transcripts.

FISH may overcome this, but it is not an automated large-scale approach. Using branched DNA (bDNA) technology, we applied single-molecule FISH (sm-FISH) to automated large-scale experiments. bDNA sm-FISH allows the use of one standard protocol and automation with high-throughput liquid-handling equipment and high-resolution screening microscopes. In conjunction with high-performance computing, bDNA sm-FISH enables the large-scale multivariate profiling of RNA transcript abundance as well as subcellular localization and patterning in thousands of single human cells per transcript with single-molecule sensitivity.

RESULTS

bDNA allows accurate single-molecule RNA measurements

In bDNA FISH, for which reagents are available from Advanced Cell Diagnostics and Affymetrix, multiple pairs of primary probes hybridize to two consecutive regions of 20–30 nucleotides at multiple positions along the transcript. Each primary probe pair jointly provides a hybridization site for a preamplifier probe, which hybridizes multiple amplifier probes that allow binding of a large number of labeled probes[10–14] (**Fig. 1a**). This contrasts with the most widely used sm-FISH approach, o-nuc sm-FISH, which employs oligonucleotides labeled with 1–5 fluorophores and lacks a signal-amplification step[15,16] (**Fig. 1b**). Consequently, o-nuc sm-FISH required a 600-times-longer exposure and a 100-times-greater camera gain than bDNA FISH to generate images with discernible spots for endogenous *MYC* mRNA in HeLa cells using a 100×/1.49–numerical aperture (NA) oil-immersion objective and electron-multiplying charge-coupled device (EMCCD) cameras (**Fig. 1c**,**d** and **Supplementary Fig. 1a**,**b**). With these different settings for exposure and gain, both approaches resulted in similar spot counts: 191.0 ± 66.4 (mean ± s.d.) spots per cell for bDNA FISH (n = 28 cells) and 189.0 ± 61.0 spots per cell for o-nuc sm-FISH (n = 20 cells). Under equal imaging conditions, bDNA spots were 100 times brighter than o-nuc spots (**Fig. 1e–g** and **Supplementary Fig. 1**), resulting in a signal-to-noise ratio that was at least 2–3 times higher than that of o-nuc sm-FISH (**Fig. 1h** and **Supplementary Note 1**). Furthermore, by labeling the same transcript with two different probe set types (**Supplementary Fig. 2a–c** and **Supplementary Note 2**), 80.8% (n = 4,703 spots) of *KIF11* transcripts and 84.58% (n = 2,979 spots) of *ERBB2* transcripts labeled with type 1 probe sets were also labeled with type 6 probe sets, which are similar accuracies to that reported for o-nuc sm-FISH[16]. Thus, bDNA FISH and o-nuc sm-FISH detected comparable numbers of discrete spots in single cells with a similar accuracy, but bDNA FISH yielded brighter spots with a better signal-to-noise ratio.

bDNA sm-FISH allows high-throughput RNA measurements

We next used a fully automated confocal microscope to image large fields of cells with a 40×/0.95-NA air objective and scientific

[1]Faculty of Sciences, Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [2]Systems Biology PhD program, Life Science Zurich Graduate School, ETH Zurich and University of Zurich, Zurich, Switzerland. [3]These authors contributed equally to this work. Correspondence should be addressed to L.P. (lucas.pelkmans@imls.uzh.ch).

RECEIVED 21 MARCH; ACCEPTED 28 AUGUST; PUBLISHED ONLINE 6 OCTOBER 2013; DOI:10.1038/NMETH.2657

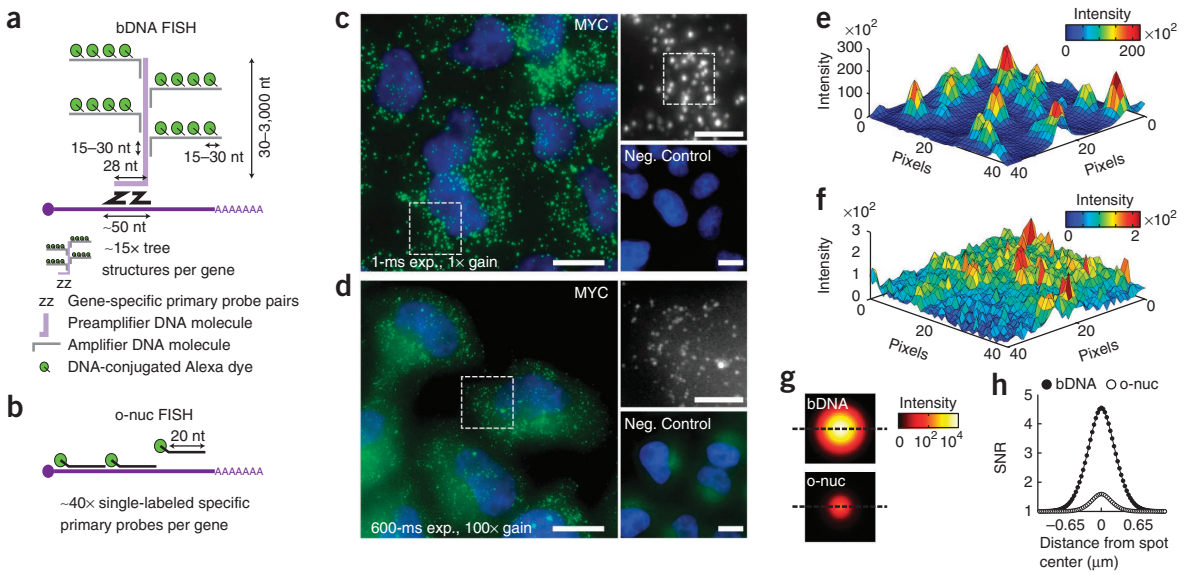


Figure 1 | bDNA FISH results in bright spots with high signal-to-noise ratio (SNR). (**a**) The bDNA FISH technique. Gene-specific primary probe pairs hybridize to the targeted RNA; tree-like structures composed of preamplifiers, amplifiers and labeled probes can be built onto these pairs, leading to signal amplification. nt, nucleotides. (**b**) The o-nuc FISH technique. The primary probes are directly labeled with a single fluorophore. (**c**) sm-FISH of endogenous MYC in HeLa cells with the bDNA method (green). Images were taken on an epifluorescence microscope using a 100×-magnification oil-immersion objective (NA = 1.49) and a back-illuminated EMCCD camera. The negative control with no primary probe pairs is also shown (bottom right). Cell nuclei are stained with DAPI (blue). Scale bars (**c**,**d**), 13 μm (overview images) and 5 μm (insets). (**d**) As in **c** but using o-nuc sm-FISH. (**e**) Intensity profile of the marked region in the top right subpanel of **c** after extracellular background subtraction. (**f**) As in **e** but for the area marked in **Supplementary Figure 1c**, the settings for which were an exposure time of 1 ms and a camera gain of 1. (**g**) Mean-modeled spots at subpixel resolution for bDNA sm-FISH and o-nuc sm-FISH after local background subtraction using a 1-ms exposure time and camera gain set to 1 (n = 100 detected spots). Dashed lines mark the spot equator. (**h**) SNR (**Supplementary Note 1**) along the equator line of the modeled subpixel spots after extracellular background subtraction; n = 100 detected spots.

complementary metal-oxide semiconductor (sCMOS) cameras. We performed FISH against the endogenous transcripts of *ERBB2*, *MYC* and *TFRC* in ~10^4 HeLa cells per gene in a 384-well plate format. Spots could be observed for each gene with the bDNA method only (**Supplementary Fig. 3**), and this method generated a highly reproducible mean number of spots per cell (23.41 ± 0.47, 203.01 ± 8.02 and 187.93 ± 6.88 for *ERBB2*, *MYC* and *TFRC*, respectively; n = 4 wells; **Supplementary Table 1**). Notably, the median number of spots per cell detected for *MYC* was comparable to that obtained with bDNA (P = 0.54, Mann-Whitney-Wilcoxon test) and o-nuc sm-FISH (P = 0.52, Mann-Whitney-Wilcoxon test) using 100×/1.49-NA magnification and EMCCD cameras.

To confirm that the spots were specific for *ERBB2*, *MYC* and *TFRC*, we performed gene silencing with RNAi. The spot-count reduction observed was strong and comparable to that determined from qPCR measurements (**Supplementary Fig. 3** and **Supplementary Table 1**). Furthermore, probe pairs against the *Escherichia coli* gene *dapB* showed a false positive rate of 0.44 ± 1.0 mean spots per cell (n = 21,094 cells). To test nuclear accessibility of the bDNA probes, we performed bDNA FISH against the nuclear-localized *SNORD3* transcripts and found no signal in the nucleus (**Supplementary Fig. 4a**,**b**). Although acetic acid in the fixation buffer[17] increased the nuclear signal for *SNORD3* and *HPRT1*, it reduced cytoplasmic spots (**Supplementary Fig. 4b**,**c**) leading to inaccurate measurements of the mature mRNA for *HPRT1* (ref. 18).

Next we tested the number of primary probe pairs that ensures that each transcript in the cytoplasm is detected by the signal of at least one primary probe pair. For both *ERBB2* and *HPRT1*, ten primary probe pairs allowed a detection of more than 80%, and 15 primary probe pairs allowed a detection of more than 90%, of the maximum number of detectable transcripts (**Supplementary Fig. 5a**,**b**). Single-cell distributions of spots per cell and their Fano factors, i.e., variance divided by mean spots per cell, also stabilized from ten primary probe pairs onwards (**Supplementary Fig. 5c**–**e**).

We then evaluated the single-spot detection accuracy of high-throughput bDNA FISH in single cells (**Supplementary Fig. 6**). The single-cell correlations of spot counts per cell for *KIF11* and *ERBB2* transcripts labeled simultaneously with two probe sets of different color (**Supplementary Fig. 3a**) were 0.976 and 0.836, respectively (Pearson correlation; **Supplementary Fig. 6a**,**b**). We estimated that for *KIF11*, 2.5% of the total cell-to-cell variability was of technical origin, whereas for *ERBB2* this was 21.8% (**Supplementary Fig. 6**). The higher fraction of technical variance in single-cell measurements for *ERBB2* was likely due to its lower expression (24.16 ± 14.55 spots per cell, n = 10,524 cells) compared to *KIF11* (73.23 ± 52.01 spots per cell, n = 10,223 cells). Thus, bDNA FISH with 15 primary probe pairs is suitable for sensitive, specific and reproducible high-throughput transcript quantification in 384-well plates at single-molecule and single-cell resolution for both low-and high-abundance transcripts.

Experimental and image-analysis pipeline

To assess the feasibility of applying our approach at the genome scale, we constructed a library of bDNA probes in 384-well

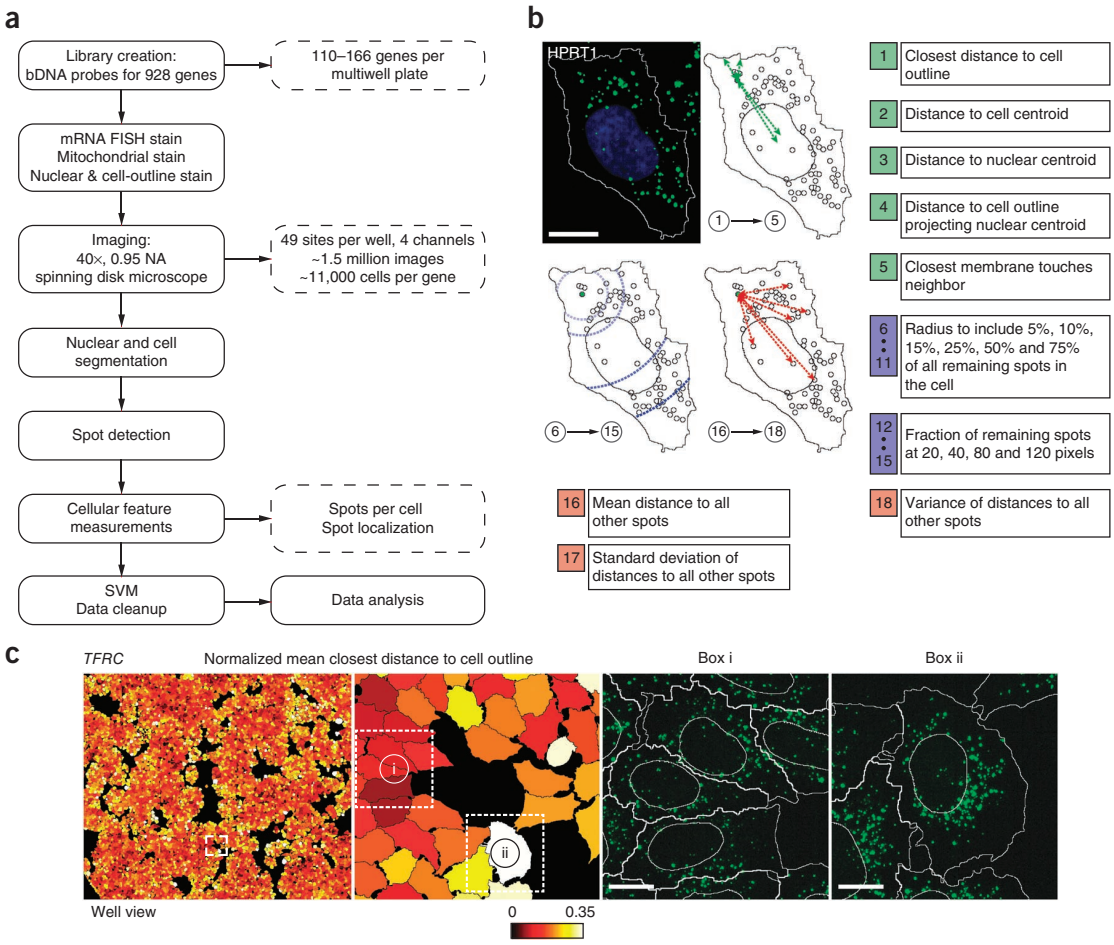


Figure 2 | Image-based transcriptomics pipeline. (**a**) Primary probes for 928 genes were plated within the center 180 wells of 384-well plates. 384-well plates containing cells were then stained in parallel with the bDNA sm-FISH reagents, MitoTracker to stain mitochondria, DAPI to stain nuclei and a protein-reactive fluorescent dye to stain whole cells. Plates were imaged at 40× magnification. Images were analyzed using CellProfiler and a custom spot-detection algorithm. Supervised machine learning (SVM) was applied to ensure high data quality by eliminating undesired phenotypes and segmentation and staining artifacts. (**b**) Features extracted to describe spot localization in single cells. Features 1–5 map every spot with respect to the cell and the nucleus. Features 6–18 map a spot relative to all other spots in the cell. (**c**) Mean closest distance to the cell outline (divided by the square root of the cell area in pixels) of all spots in a cell for *TFRC* transcripts in a population of cells. Green, bDNA sm-FISH; blue, DAPI (cell nucleus). Scale bars, 13 µm.

plates targeting 928 human genes involved in basic cellular functions, cancer, signaling, endocytosis and metabolism (**Supplementary Table 2**). In addition, we modified existing algorithms[16,19–21] to create a robust high-throughput spot-detection pipeline (**Supplementary Note 3**, **Supplementary Fig. 7** and **Supplementary Software**). We automated the experimental protocol using a liquid-handling platform (**Supplementary Protocol**), and image analysis[19] and supervised machine learning data cleanup[22] were submitted to high-performance computing using iBRAIN[23]. As a proof of principle, we performed two independent biological replicates of *in situ* transcriptomics in an unperturbed HeLa cell line (**Fig. 2a**).

We acquired confocal images in ten *z* planes, with a step size of 1 µm, covering the full cellular height at 49 sites in each well. Because two-dimensional spot detection on maximum-intensity projections of *z* stacks yielded virtually identical numbers of spots per cell as three-dimensional spot detection (**Supplementary Fig. 7i**), we performed all our quantifications on projected *z* stacks. We obtained 18 primary spot features that reflect the relative localization of each spot in a single cell, with respect to both the cell and other spots (**Fig. 2b**,**c**). To give an impression of the information contained in one such feature, we depicted the single-cell values for mean closest distance of detected spots to the cell outline for the transcript *TFRC* (transferrin receptor 1) in a segmented population of cells (**Fig. 2c**).

High-throughput quantitative image-based transcriptomics

The mean number of spots per cell for each gene was highly reproducible between the two biological replicates (Pearson correlation of 0.989; **Fig. 3a** and **Supplementary Table 3**). In addition, the absolute gene expression level of control genes across plates was very similar (**Supplementary Fig. 8**). When comparing distributions of single-cell spot counts of each gene with the negative control *dapB* (**Fig. 3b**), we found that 857 of 928 gene transcripts contained a significant fraction of cells with spot counts higher than those for *dapB* ($P < 10^{-4}$ for both replicates; **Supplementary Note 4** and **Fig. 3c**). These 857 detected genes displayed a bimodal distribution of low- and high-expressed transcripts with

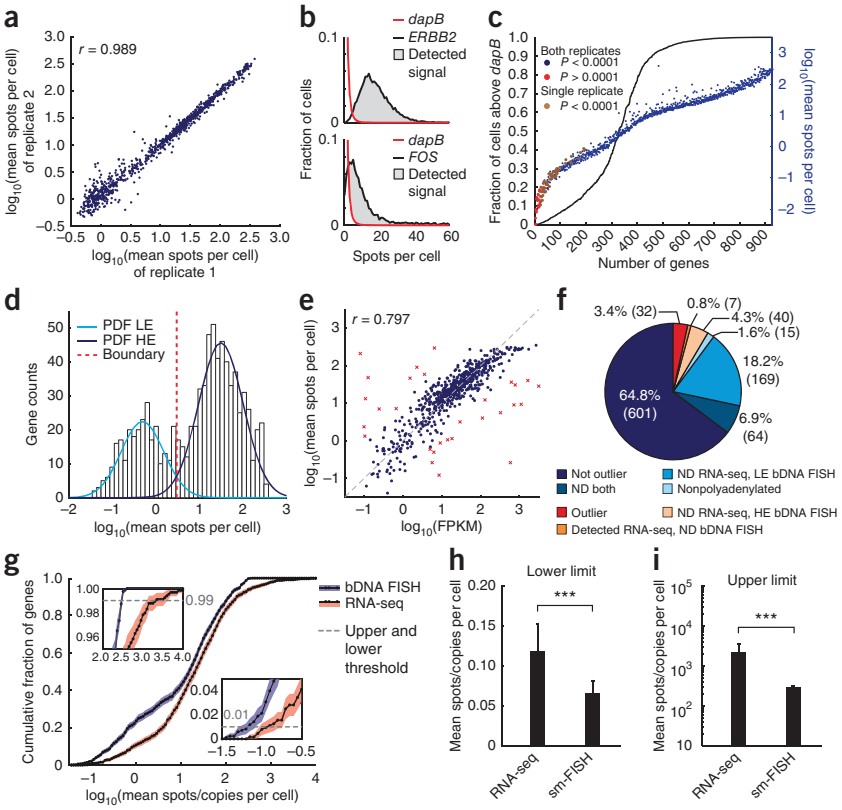


Figure 3 | Image-based transcriptomics is reproducible, sensitive and comparable to RNA-seq. (**a**) $\log_{10}$(mean spots per cell) correlation of biological replicates. The Pearson correlation is shown. (**b**) Relative distribution of *dapB* compared to those of two examples, *ERBB2* and *FOS* (replicate 1). The gray area is the fraction of cells above background (or detected signal) for a given gene. (**c**) Fraction of cells above background (black line) and corrected mean expression level (data points) in $\log_{10}$(mean spots per cell) for each gene; $n = 500$ bootstraps. Colors indicate whether the fraction of cells above background reached significance ($P < 1.0 \times 10^{-4}$) in none (red), one (brown) or both replicates (blue). (**d**) Distribution of corrected $\log_{10}$(mean spots per cell) for blue data points in **c** (857 genes). Solid lines indicate the probability density function (PDF) of low-expressed (LE) and high-expressed (HE) genes. The dashed line is the estimated boundary between the two classes (3.01 spots per cell). (**e**) Correlation of RNA-seq, $\log_{10}$(fragments per kilobase of exon model per million mapped reads (FPKM)), and high-throughput bDNA sm-FISH $\log_{10}$(mean spots per cell). Outliers are shown in red. *r* is the Pearson correlation before outlier rejection. The dashed line is a guide for the eye. (**f**) Detailed comparison of transcript detection for RNA-seq and high-throughput bDNA sm-FISH. ND, not detected. (**g**) Cumulative fraction of genes as a function of the expression level in $\log_{10}$(mean spots/copies per cell) for RNA-seq and high-throughput bDNA sm-FISH. Smaller panels show the estimation of lower and upper limits of detection (dashed lines). Shaded areas represent the s.d. of 1,000 bootstraps. (**h**,**i**) Quantification of the lower (**h**) and upper (**i**) limits of detection for RNA-seq and high-throughput bDNA sm-FISH. Error bars, s.d. of 1,000 bootstraps. $***P < 1.0 \times 10^{-3}$.

a boundary between them at 3.01 ± 0.50 mean spots per cell ($n = 1{,}000$ bootstrapped samples; **Fig. 3d**, **Supplementary Fig. 9** and **Supplementary Note 4**). Such a boundary was previously estimated at a lower value[24].

Notably, the correlation between mean spot count per cell and transcript abundance measured with RNA-seq (**Supplementary Fig. 10**) was 0.797 (Pearson correlation) or 0.842 (Spearman correlation) (**Fig. 3e**), which increased to 0.917 (Pearson correlation) or 0.915 (Spearman correlation) after outlier rejection. For 71.7% of transcripts, both methods either detected a signal at similar levels (64.8%) or did not detect a signal (6.9%; **Fig. 3f**). 22.5% of transcripts were detected only by bDNA sm-FISH (18.2% as low-expressed transcripts), whereas 0.8% of transcripts were detected only by RNA-seq (0.54% as low-expressed transcripts, not shown).

Of the remaining 5% of transcripts, 1.6% were detected only by bDNA sm-FISH because they were nonpolyadenylated, whereas 3.4% were detected by both methods but their levels did not correlate (**Fig. 3e**). Comparing the detection sensitivity and dynamic range of high-throughput bDNA sm-FISH with RNA-seq revealed that at the lower limit of detection, high-throughput bDNA sm-FISH was more sensitive than RNA-seq (0.066 ± 0.015 and 0.118 ± 0.034 spots/copies per cell, respectively; $n = 1{,}000$ bootstrapped samples; **Fig. 3g**,**h**). At the upper limit of detection, high-throughput bDNA sm-FISH showed a ceiling effect at 288.98 ± 18.24 spots/copies per cell at the mean level ($n = 1{,}000$ bootstrapped samples; **Fig. 3g**,**i**). At the single-cell level, however, we obtained spot counts higher than 1,500 (for example, for 18S1–18S5 RNA, *CYTB* and *GAPDH*; not shown). For RNA-seq, the upper limit of detection for the genes in our library was 2,262.41 ± 1,278.74 copies per cell ($n = 1{,}000$ bootstrapped samples; **Fig. 3g**,**h**).

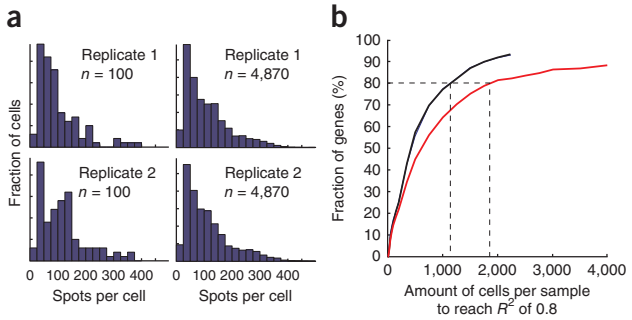


Figure 4 | Minimum number of cells required for reproducible single-cell distributions of transcript abundance. (**a**) Example distributions of transcript abundance of the cell cycle–associated gene *PLK1* in two biological replicate measurements using a sample size of 100 single cells or a sample of 4,870 single cells. If only 100 cells are sampled, the distributions of single-cell spot counts (at a bin size of 25 spots) are dissimilar. (**b**) Number of cells that need to be sampled to reach a coefficient of determination (R^2) of 0.8 between single-cell spot count distributions within (black and blue lines) or between (red line) the two replicates. Dashed lines indicate requirement of cells for 80% of all genes.

Thus, high-throughput bDNA sm-FISH generates highly reproducible results and is a quantitative method for large-scale transcriptomics with high sensitivity that rivals RNA-seq for low-abundance transcripts.

Requirements for reproducible single-cell distributions

Most studies on cell-to-cell variability in RNA transcript copy number have so far relied on the quantification of, at maximum, several hundred single cells[24–27]. However, it is unclear how many

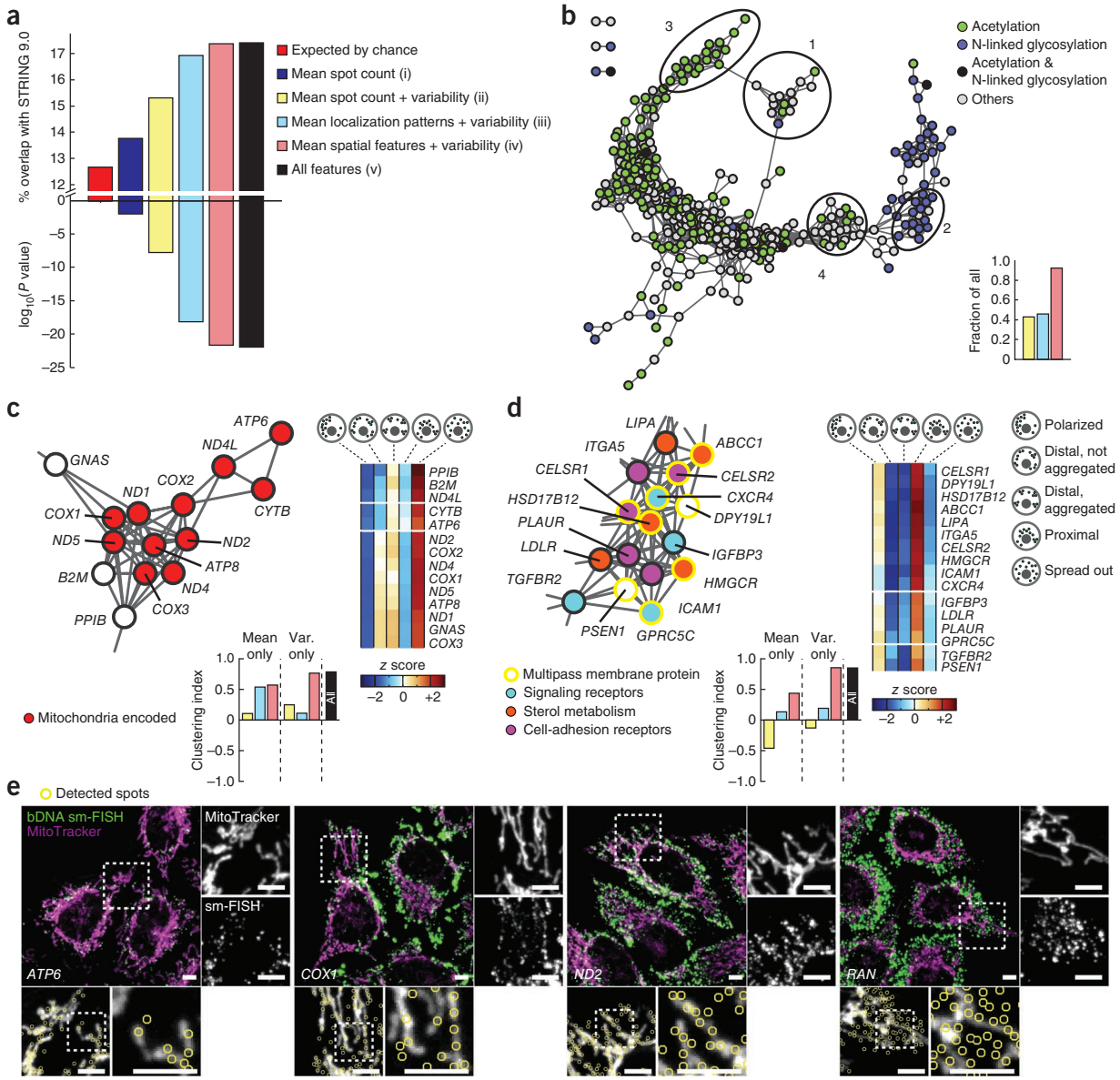


Figure 5 | Quantitative signatures of the *in situ* transcriptome. (**a**) Overlap of the 5% smallest pairwise gene-gene distances with known gene interactions in STRING 9.0 and their respective *P* values. Data are shown for five different sets of features: (i) mean RNA spot count per cell (blue); (ii) mean RNA spot count per cell and features of its distribution (variability) (yellow); (iii) mean single-cell classification of localization patterns per gene and features of the classification distributions (light blue; see **Supplementary Fig. 12**); (iv) mean spatial features of spots per gene and features of their distributions (salmon); and (v) the combination of all extracted features (black). See also **Supplementary Figure 14b.** (**b**) Gene network (4,873 edges) obtained with the 5% smallest gene-gene distances derived from the combination of all features (black bar in **a**). Only connected genes are shown (96.8% of included genes). Node colors indicate genes encoding acetylated proteins (green), N-linked glycosylated proteins (blue), those that undergo both modifications (black) and others (gray). The bar graph indicates the fraction of edges that are also retrieved with three specific feature subsets, subsets ii–iv in **a**; color-coding as in **a**. Subregions in the network correspond to **c** (subregion 1), **d** (subregion 2) and **Supplementary Figure 15d**,**e** (subregions 3 and 4). (**c**,**d**) Subregion 1, a tight cluster of genes encoded in the mitochondrial genome (red nodes, **c**); and subregion 2, a tight cluster of genes encoding cell-adhesion receptors (purple nodes, **d**), signaling receptors (light blue nodes, **d**) or proteins involved in sterol metabolism (orange nodes, **d**). Subregion 2 contains multipass membrane proteins (yellow-outlined nodes, **d**). *z*-scored mean classification distributions of cells for all five main types of single-cell spot localization patterns (specified at right) are shown as clustered heat maps. Bar graphs indicate the clustering index for three specific feature subsets, subsets ii–iv in **a**; color-coding as in **a**. (**e**) Subcellular localization of transcripts from the mitochondria-encoded genes *ATP6*, *COX1* and *ND2*, as well as of the transcripts from *RAN* (which does not cluster in subregion 1), with respect to MitoTracker. Yellow circles are detected spots. Scale bars, 5 µm.

cells must be sampled to obtain reproducible single-cell spot count distributions. We therefore compared random samples of increasing number of single cells for each gene to a second sample from the same cell population and a sample derived from the biological replicate (**Fig. 4a** and **Supplementary Fig. 11a**). Across all tested genes, 100 cells sufficed to obtain reproducible measurements of the mean, variance and Fano factor (Pearson correlation of 0.997, 0.951 and 0.910, respectively; **Supplementary Fig. 11b**). However, the third, fourth and fifth central moments required 215, 274 and 1,764 single cells, respectively, to obtain a Pearson correlation of 0.75 (**Supplementary Fig. 11c**). Furthermore, correlating whole spot count distributions revealed that at least 1,100 single cells were required to reach a high coefficient of determination ($R^2 = 0.8$) for 80% of the genes when different samples from the same cell population were compared, and 1,800 cells were required for samples coming from different biological replicates (**Fig. 4b** and **Supplementary Fig. 11d–g**). Thus, for most genes in a nonsynchronized unperturbed HeLa cell line, at least 1,000 single cells must be sampled to obtain reproducible single-cell distributions of transcript copy number.

Quantitative signatures of the *in situ* transcriptome

Finally, we wrote algorithms to harness the multivariate feature set quantifying subcellular localization and patterning of single transcripts within thousands of single cells. The first algorithm performs unsupervised clustering of all single cells to identify the main types of subcellular spot localization patterns, aiding biological interpretability (**Supplementary Fig. 12**, **Supplementary Note 5** and **Supplementary Software**). In the generated data set, this algorithm revealed five main types of single-cell patterns: a polarized distribution, distal distribution, distal and aggregated distribution, proximal (perinuclear) distribution and spread-out distribution of spots (**Supplementary Fig. 13**). The second algorithm maximizes the information contained within the multivariate feature set by computing additional features describing the variability of the spot count per cell and the spatial distribution of spots (**Supplementary Fig. 14**, **Supplementary Table 4** and **Supplementary Note 6**).

We then tested various combinations of the information obtained from these two algorithms to evaluate their ability to cluster genes that are functionally associated in a database of known and predicted protein interactions (STRING v.9.0; ref. 28) (**Supplementary Fig. 14**). This analysis revealed that quantitative information about subcellular patterns and spatial properties of transcripts and their variability across single cells were more powerful at identifying functional interactions than were features of mean spot count and its variability (**Fig. 5a** and **Supplementary Fig. 15a**,**b**).

We next created a network (**Fig. 5b**) from the top 5% of calculated gene-gene distances on the basis of their similarity in transcript features. The majority of edges in this network could be derived from spatial features and their variability (**Fig. 5b** and **Supplementary Fig. 15c**). Globally, genes that encode acetylated proteins translated in the cytosol separated from genes that encode N-linked glycosylated proteins translated at the endoplasmic reticulum (ER). Extensive subclustering within these two domains indicated that our feature set revealed details of subcellular patterning of transcripts and its cell-to-cell variability with functional relevance beyond general differences in translation sites.

A specific isolated region in the network (**Fig. 5b**) was formed by a tight subcluster of 11 of the 13 mRNA-coding genes encoded by mitochondria and showed an enrichment of cells with a spread-out and a distal distribution of transcripts (**Fig. 5c**). This subcluster was distinguished from its immediate surrounding by features of subcellular patterning as well as of spatial properties and their variability (**Fig. 5c**), whereas features of transcript abundance and variability did not contribute to this subclustering. Further analysis revealed that whereas transcripts of *ATP6*, *COX1* and *ND2* localized to stained mitochondria, transcripts of *RAN*, which is not part of this cluster (although it was nearby in the network), fell outside of the stained mitochondria (**Fig. 5e**).

The region of the network consisting of genes encoding N-linked glycosylated proteins also showed subclustering. One of these subclusters (**Fig. 5d**) consisted of genes encoding proteins involved in sterol metabolism and cell adhesion, and signaling receptors. The majority of these were multipass membrane proteins. This subcluster displayed an enrichment for cells with a perinuclear distribution of transcripts (**Fig. 5d**), a result consistent with localization to the ER[28]. The subcluster was distinguished from its immediate surroundings in the network by spatial properties and their variability, suggesting localization at specific subdomains of the ER. Features of transcript abundance alone would prevent this subclustering (**Fig. 5d**). Also, the region in the network enriched for acetylated proteins displayed further subclustering (**Fig. 5b**), with one subcluster of genes encoding ribosomal proteins and proteins involved in glycolysis and energy production and another subcluster of genes encoding proteins involved in endocytosis and ubiquitination (**Supplementary Fig. 15d**,**e**).

Taken together, the extracted feature set contains multiple types of information about specific *in situ* signatures of the transcriptome. In particular, features of subcellular localization and patterning and their variability allow the unbiased identification of functional interactions between genes without the need for any perturbation or costaining.

DISCUSSION

We have demonstrated the feasibility of large-scale image-based transcriptomics by applying sm-FISH in an automated high-throughput manner in human tissue culture cells, achieving comparable results to RNA-seq at the mean expression level. Most of the bDNA sm-FISH reagents used in this study were produced by Affymetrix upon our request and have since become available to other customers, thereby making our approach readily accessible. Currently, bDNA sm-FISH shows limited detection of nuclear transcripts, has less dynamic range than RNA-seq for high-abundance transcripts and may, for a few transcripts, obtain aberrant readouts. Another limitation is the small number of different transcripts that can be quantified in the same single cell compared to that by single-cell RNA-seq, which can achieve quantification of more than 6,000 transcripts per cell[8]. However, the bDNA signal amplification tree may allow extensive barcoding, which could be exploited for single-cell multiplexing in the near future[29–31].

High-throughput bDNA sm-FISH scales dramatically better than single-cell RNA-seq in the number of single cells that can be measured within the same sample[8], which is important for reproducible measurements of cell-to-cell variability in RNA transcript

abundance. It is also more sensitive than single-cell RNA-seq for low-abundance transcripts, reveals absolute copy numbers and allows the quantification of multivariate features of transcript patterning within and across thousands of single cells. Our analysis of these features revealed that shared properties of the variability in subcellular transcript localization across unperturbed single cells outperform cell-to-cell variability in transcript abundance in retrieving functional associations between genes.

Further development in the types of analysis shown here combined with perturbation experiments will increase the power of this approach. We expect that high-throughput bDNA sm-FISH will find broad applications as it can be directly included in various image-based approaches. This will enable a more direct examination of the causal links between molecular and phenotypic cell-to-cell variability.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to acknowledge B. Snijder and Y. Yakimovich for help with computational analysis and infrastructure, J. Patterson for assistance, Q. Nguyen and S. Lai from Affymetrix for helpful comments on experimental procedures, J. Ellenberg (European Molecular Biology Laboratory) for reagents, and all members of the lab for useful comments on the manuscript. L.P. acknowledges financial support for this project from SystemsX.ch, the European Union, University of Zurich and University of Zurich Research Priority Program in Systems Biology and Functional Genomics.

AUTHOR CONTRIBUTIONS

L.P. initiated the study. N.B., T.S. and L.P. designed and analyzed the experiments and wrote the manuscript. N.B. and T.S. performed the experiments.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* **21**, 33–37 (1999).
2. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
3. Wilhelm, B.T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239–1243 (2008).
4. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
5. Tang, F. *et al.* mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
6. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
7. Ramsköld, D. *et al.* Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
8. Shalek, A.K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
9. Goetz, J.J. & Trimarchi, J.M. Transcriptome sequencing of single cells with Smart-Seq. *Nat. Biotechnol.* **30**, 763–765 (2012).
10. Lau, J.Y. *et al.* Significance of serum hepatitis C virus RNA levels in chronic hepatitis C. *Lancet* **341**, 1501–1504 (1993).
11. Kern, D. *et al.* An enhanced-sensitivity branched-DNA assay for quantification of human immunodeficiency virus type 1 RNA in plasma. *J. Clin. Microbiol.* **34**, 3196–3202 (1996).
12. Player, A.N., Shen, L.P., Kenny, D., Antao, V.P. & Kolberg, J.A. Single-copy gene detection using branched DNA (bDNA) in situ hybridization. *J. Histochem. Cytochem.* **49**, 603–612 (2001).
13. Kenny, D., Shen, L. & Kolberg, J.A. Detection of viral infection and gene expression in clinical tissue specimens using branched DNA (bDNA) *in situ* hybridization. *J. Histochem. Cytochem.* **50**, 1219–1227 (2002).
14. Ma, X.-J., Wu, X. & Luo, Y. Biomarkers for differentiating melanoma from benign nevus in the skin. US patent application 20120071343 (2012).
15. Femino, A.M., Fay, F.S., Fogarty, K. & Singer, R.H. Visualization of single RNA transcripts *in situ*. *Science* **280**, 585–590 (1998).
16. Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
17. Chartrand, P., Bertrand, E., Singer, R.H. & Long, R.M. Sensitive and high-resolution detection of RNA *in situ*. *Methods Enzymol.* **318**, 493–506 (2000).
18. Bhatt, D.M. *et al.* Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell* **150**, 279–290 (2012).
19. Carpenter, A.E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
20. Raj, A. & Tyagi, S. Detection of individual endogenous RNA transcripts *in situ* using multiple singly labeled probes. *Methods Enzymol.* **472**, 365–386 (2010).
21. So, L.H. *et al.* General properties of transcriptional time series in *Escherichia coli*. *Nat. Genet.* **43**, 554–560 (2011).
22. Rämö, P., Sacher, R., Snijder, B., Begemann, B. & Pelkmans, L. CellClassifier: supervised learning of cellular phenotypes. *Bioinformatics* **25**, 3028–3030 (2009).
23. Snijder, B. *et al.* Single-cell analysis of population context advances RNAi screening at multiple levels. *Mol. Syst. Biol.* **8**, 579 (2012).
24. Hebenstreit, D. *et al.* RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol. Syst. Biol.* **7**, 497 (2011).
25. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).
26. Trcek, T., Larson, D., Moldón, A., Query, C. & Singer, R. Single-molecule mRNA decay measurements reveal promoter- regulated mRNA stability in yeast. *Cell* **147**, 1484–1497 (2011).
27. Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209–1222 (2012).
28. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**, D561–D568 (2011).
29. Lubeck, E. & Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat. Methods* **9**, 743–748 (2012).
30. Nguyen, Q.N., Lipshutz, R.J. & Ma, Y. Methods of labeling cells, labeled cells, and uses thereof. US patent application 20120178081 (2012).
31. Levesque, M.J. & Raj, A. Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nat. Methods* **10**, 246–248 (2013).

ONLINE METHODS

Cell culture. HeLa Kyoto cells were kindly provided by J. Ellenberg (EMBL, Heidelberg). Cells were tested for identity by karyotyping and tested for absence of mycoplasma before use. Culturing was done in DMEM (Gibco) supplemented with 10% FCS and glutamine (complete medium). Seeding was at a density of 700 cells per well when using 384-well plates (Greiner) and 10,000 cells per well when using a LabTek chambered #1.0 borosilicate coverglass system of eight chambers. Cells were incubated for 3 d at 37 °C, 95% humidity and 5% CO_2. For image-based transcriptomics, a full cell culture was regrown from a single cell in six passages, after which cells were harvested, frozen and kept at −80 °C until use. Only cells imaged at 100× magnification were grown in LabTek chambers.

Microscopy. For high-magnification oil-immersion imaging, we used a Nikon Eclipse Ti inverted fluorescence microscope, an Apo TIRF 100× objective (Nikon) of 1.49 NA and an EMCCD camera (ImageEM 1K C9100-14, Hamamatsu). High-throughput *in situ* transcriptomics imaging, was done with an automated spinning disk microscope from Yokogawa (CellVoyager 7000), with an enhanced CSU-X1 spinning disk (Microlens-enhanced dual Nipkow disk confocal scanner, wide view type), a 40× Olympus objective of 0.95 NA, and a Neo sCMOS cameras (Andor, 2,560 × 2,160 pixels), acquiring 49 sites per well and ten *z* planes per site spanning 9 µm (**Supplementary Table 5**). The number of *z* planes was chosen so that every spot was visible in at least two planes as described previously[20]. The primary probe pair saturation curves were measured with an ImageXpress Micro fluorescence microscope (Molecular Devices), a Plan Apo 40× objective (Nikon) of 0.95 NA and a CoolSNAP HQ camera.

Oligonucleotide single-molecule RNA FISH. Quasar 570–labeled oligonucleotide Stellaris FISH RNA probes targeting *TFRC*, *MYC* and *ERBB2* mRNA were obtained from Biosearch Technologies. Probe hybridization was performed as indicated by the manufacturer.

Branched DNA single-molecule RNA FISH. All gene-specific primary probe pairs, amplification systems and custom-designed probes for measurement of saturation curves and double-labeling experiments were purchased from Affymetrix upon specific request and have since been made commercially available. Experiments were performed following the **Supplementary Protocol**. In the signal-saturation experiments, 15 individual primary probe pairs targeting *ERBB2* and *HPRT* were acquired from Affymetrix. Probe pairs were then combined in such a way to generate 30 primary probe-pair combinations per gene spanning a range of 1–15 targeted sites per gene. bDNA sm-FISH was then performed as described in the **Supplementary Protocol**. For acetic acid experiments, glacial acetic acid was added at the required [v/v]% to the fixation solution (4% paraformaldehyde in PBS).

Calculation of signal-to-noise ratios. Spot detection of 100×-magnification images was done as described in the **Supplementary Note 2**, although no spot bias correction was applied. Calculation of the signal-to-noise ratio is described in **Supplementary Note 1.**

Library construction. The final library was composed of probes against 925 human genes of general interest (**Supplementary Table 2**), probes against three positive-control genes (*ERBB2*, *HPRT* and *ACTB*) covering a wide range of expression levels and probes against a bacterial gene (*dapB*) as negative control. The library was mostly composed of QuantiGene View RNA type I primary probe pairs, although some genes were targeted with QuantiGene View RNA types VI, VIII or X. Primary probes for all genes were then organized in six 384-well plates according to plate layout in the **Supplementary Protocol**. Aliquots in such plates were diluted 1:5 and then 1:10 to arrive at the working concentration of primary probe sets.

siRNA gene knockdown. Validated siRNAs targeting *ERBB2* (SI02223571, Hs_ERBB2_14), *MYC* (SI00300902, Hs_MYC_5) and *TFRC* (SI00301896, Hs_TFRC_5) were obtained from Qiagen. Reverse transfection was done using Lipofectamine2000 (Invitrogen) according to the manufacturer's specifications. Cells were fixed 3 d after transfection for bDNA sm-FISH.

Quantitative reverse-transcription PCR. RNA was extracted with the RNeasy mini kit including the optional on-column DNA digestion (Qiagen) and reverse transcribed with oligo(dT) primers using the Transcriptor High Fidelity cDNA Synthesis kit (Roche) according to the manufacturers' protocols. Real-time PCR was done with a Mesa Green qPCR Mastermix Plus for Sybr Assay (Eurogentec) with the following primers. hs_TFRC_fwd: catttgtgagggatctgaacca; hs_TFRC_rev: cgagcagaatacagccactgtaa; hs_ERBB2_fwd: agaccatgtccgggaaaacc; hs_ERBB2_rev: caggtagc tcatccccttgg; hs_MYC_fwd: cgactctgaggaggaacaagaa; hs_MYC_rev: actctgacctttgccaggag; hs_TBP_fwd: gcccgaaacgccgaatata; hs_TBP_rev: cgtggctctcttatcctcatga; hs_EEF1A1_fwd: agcaaaaa tgacccaccaatg; hs_EEF1A1_rev: ggcctggatggttcaggata.

Image analysis. All images were analyzed with the image analysis software CellProfiler[19]. Methods required for this study were implemented in Matlab and, when possible, as new CellProfiler modules (see **Supplementary Software**). Nuclei were segmented using images from the 4,6-diamidino-2-phenylindole (DAPI) staining. The cell outlines were then identified using the watershed algorithm. Spot detection was carried out as described in **Supplementary Note 2**. Standard CellProfiler features for intensity, size and texture were then extracted for nuclei and cells. For data cleanup, we applied supervised machine learning with CellClassifier[22,23] to exclude cells showing segmentation problems or aberrant staining, undergoing mitosis or being multinucleated. Computational image analysis was done using the Brutus high-performance computing cluster (ETH Zurich) and the computational task manager iBRAIN[23]. All modules and source code developed for this project can be downloaded at https://github.com/pelkmanslab/.

RNA-seq. Total RNA was extracted from cell lysates using the RNeasy mini kit (Qiagen) with on-column digestion of DNA as specified by the manufacturer. Transcriptome sequencing was performed by LC Sciences. Briefly, RNA quality was assessed using the RNA 600 LabChip (Agilent). Sample preparation was done using the TruSeq RNA Sample Prep Kit v.2 (RS-122-2001, Illumina) as specified by the manufacturer. Enrichment for polyadenylated

doi:10.1038/nmeth.2657

RNA was done using poly(T) beads, and cDNA was obtained from random primers after RNA fragmentation. Sequencing was done using a HiSeq 2000 sequencer from Illumina. Read alignment was done using Bowtie v.0.12.7 (ref. 32) against the human genome (hg19), and FPKM values were generated using TopHat v.1.3.2 (ref. 33) and Cufflinks v.1.3.0 (ref. 34). The FPKM value for a given gene was derived by adding all FPKM values assigned to all transcripts of the gene (**Supplementary Table 6**). For both RNA-seq replicates we obtained ~1.1×10^8 mappable reads, of which ~1.01×10^8 were mapped to exons, ~5.6×10^7 reads mapped to spanning exons and ~8.8×10^6 reads mapped to introns.

Estimation of boundary between low- and high-expressed transcripts. A Gaussian mixture model of corrected mean spots per cell was learned assuming two distributions representing the low- and high-expressed transcripts, respectively. Modeling was done using Matlab. The boundary between the two distributions was set where the probability of being low expressed or high expressed given a mean spot number per cell was equal, i.e., $P(w_1|x) = P(w_2|x)$, where w_1 and w_2 are the low- and high-expressed gene classes, respectively, and x represents a given mean spot number per cell. The computation of the boundary was bootstrapped 1,000 times with replacement. Then the mean boundary value and its s.d. were calculated.

Outlier detection in bDNA sm-FISH vs. RNA-seq comparison. Calculation of the fraction of cells with spot counts above background and mean spot per cell correction was performed according to **Supplementary Note 4**. The correlation plot obtained from $\log_{10}$(FPKM), and corrected $\log_{10}$(spots per cell) was regressed using robust LOESS with the Computational Statistics Matlab library[35]. The shortest distance to the regression line was measured for every gene, and outliers were defined as those points whose distance was bigger than two times the s.d. of all distances.

Calculations of upper and lower detection limits. Conversion of $\log_{10}$(FPKM) to $\log_{10}$(spots per cell) was done by linear regression and extrapolation with the 601 genes whose expression agreed between RNA-seq and high-throughput bDNA sm-FISH. Regression was done with the Matlab Statistics Toolbox "regress" function. Cumulative fractions were calculated by 1,000 bootstrap random samples of 301 genes without replacement, and upper and lower limits of detection were set to the 0.99 and 0.01 cumulative fractions. P values were calculated using a two-sample t-test.

Estimation of the minimal amount of cells required for reproducible cell-to-cell variability. For those transcripts whose mean spot count per cell agreed well with RNA-seq with uncorrected spot counts (612 genes, not shown), we randomly sampled an increasing equal number of cells from each of the two biological replicate experiments. We then calculated the distribution of single cells in each of the samples from zero spots per cell to the highest number of spots per cell, using a bin size of one spot. The Pearson correlation between two distributions within a replicate or between the two replicates was then measured. The procedure was bootstrapped 100 times, and correlation values were computed for every gene and every sampling size, from which the R^2 was then computed. We calculated the Pearson correlation of distribution mean, variance, Fano factor and central moments over all genes at each sampling size for two distributions sampled (i) within a replicate or (ii) between the two replicates. The procedure was bootstrapped 100 times.

Estimation of percentage of genes with highly reproducible multivariate transcript readouts. For each multivariate readout of each gene, its mean ranked distance to its replicate was obtained. This was done by comparing the Euclidean distance of a given gene to all genes of the replicate assay in a given feature space. The ranked distance to the replicate of the same gene was determined. To account for each gene being tested twice, we used for each the mean distance of each replicate. A readout of a gene was defined as highly reproducible when its replicate readout was within the 5% closest distances, unless otherwise specified.

Network construction from the 5% smallest gene-gene distances. Feature selection as described in **Supplementary Note 6** was performed for genes whose mean raw spot count in both replicate assays was at least ten spots per cell and whose raw spot count correlated well with RNA-seq counts (442 genes). For each set of features from individual repetitions of feature selection, Euclidean distances between genes were calculated on features normalized by z-scoring over all genes included in the feature selection. To account for differences in the total amount of features and, thus, the absolute Euclidean distance between individual rounds of elimination, we normalized the Euclidean distances by taking the square root of the square of the Euclidean distances divided by the number of features. The normalized distances at this point were averaged over all 60 iterations for each starting feature set to obtain a mean dissimilarity matrix. Next, gene-gene distances were defined as the Euclidean distance between genes using the mean dissimilarity matrix and then ranked with the smallest distance obtaining rank 1 while excluding similarities of a gene to itself. For network analysis, the top 5% ranking gene-gene distances for every feature set were used.

Calculation of the clustering index for network nodes in subregions. Networks built from different feature sets after selection were used for the calculation of the clustering index between nodes G (genes) belonging to the four subregions of interest (**Fig. 5b**). For every network, edges connecting the G nodes were defined in two categories: k edges that connect G nodes to other G nodes, i.e., these edges connect genes that are within a given subregion, or q edges that connect G nodes to other nodes in the network, i.e., genes that are outside the given subregion. Then the clustering index I is given by the following expression:

$$I = \frac{\sum_{i=1}^{n_g} \frac{(K_i - Q_i)\,\mathrm{sgn}(K_i)}{K_i + Q_i}}{n_g}$$

where n_g is the number of G nodes in the given subregion, K_i is the number of k edges connecting a given G_i node, Q_i is the number of q edges connecting a given G_i node, and $\mathrm{sgn}(K_i)$ is a sign function whose value is 0 when $K_i = 0$ and 1 when $K_i > 0$. Thus the clustering index will be positive if the given G nodes have on average more k edges than q edges.

Calculations of enrichments in STRING 9.0. The reference data set was obtained from the STRING 9.0 database (http://string-db.org/). For each gene-gene distance, the presence of a reported interaction in STRING was determined. For every feature set, the overlap was defined as the fraction of gene-gene distances that was present in STRING 9.0 and whose distance was smaller than the indicated distance. *P* values are given by the hypergeometric probability density function and are the sum of the *P* values of all possibilities that yield at least the observed amount of overlapping gene-gene interactions.

Network analysis. Network analysis and automated force-directed visualization was performed using Cytoscape[36]. Heat maps displaying clustered fractions of cells of the five main types of single-cell spot localization patterns for the example network subregions in **Figure 5c**,**d** and **Supplementary Figure 15d**,**e** were derived from the *z*-scored means of the classification distributions for every pattern type. (**Supplementary Note 6**). Hierarchical clustering using a Euclidean distance and average linkage was performed in Matlab.

Statistical analysis. The bootstrapped samples obtained from calculation of fraction of cells above background (**Supplementary Note 4**) for every replicate gene was compared to the distributions of fractions expected by random using the Mann-Whitney-Wilcoxon test implemented in Matlab. The *P* values obtained were corrected for multiple testing using the Holm-Bonferroni method. To identify genes with fractions of cells above background, we set a conservative significance value of $P = 10^{-4}$.

32. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
33. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
34. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
35. Martinez, W.L. & Martinez, A.R. *Computational Statistics Handbook with MATLAB* 2nd edn. (CRC Press, 2008).
36. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

doi:10.1038/nmeth.2657

Supplementary Information for

Large-scale image-based transcriptomics in thousands of single human cells at single-molecule resolution

Nico Battich[1,2,3], Thomas Stoeger[1,2,3], Lucas Pelkmans[1]

[1]Faculty of Sciences, Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland.

[2]Life Science Zurich Graduate School, Ph.D. program in Systems Biology.

[3] contributed equally

Correspondence should be addressed to L.P. (lucas.pelkmans@imls.uzh.ch)

Contents

Nature Methods: doi:10.1038/nmeth.2657

Nature Methods: doi:10.1038/nmeth.2657

Supplementary Protocol: High-throughput image-based transcriptomics using bDNA sm-FISH

Cells were seeded as described in Online Methods in complete medium supplemented with Penicillin/Streptomycin (Gibco). Cells were fixed in 4% paraformaldehyde for 30 min at RT. Incubation of primary probe pairs was done for 3 hrs, while for pre-amplifiers, amplifiers and label probes for 1 hr each. All hybridization steps were done at 40ºC in a Liconics rotating incubator to avoid plate positional effects. After hybridization reactions, cells were stained for 10 min with 0.2µg/ml DAPI in PBS and then incubated for 5 min in 1ng/µl of Alexa Fluor® 647 carboxylic acid, succinimidyl ester (Invitrogen) in carbonate buffer (1.95ml of 0.5M NaHCO3, 50µl of 0.5M Na2CO3 in and 8ml of water for 10ml of buffer) for detection of the cell outline. For image-based transcriptomics, prior to fixation cells were incubated in 0.5µM MitoTracker® Red CMXRos for 45 min at 37ºC for detection of mitochondria. All steps in the protocol were automated using the EL406 washer-dispenser from BioTek and a Bravo liquid handling platform from Agilent Technologies.

High-throughput in situ RNA hybridization: Main protocol

Experiment Title:
Date:

Notes

1	This protocol assumes processing of three 384-well plates in sequential order.
2	Every aspiration step was done in the EL406 BioTek washer-dispenser so that 15ul residual volume were left in the well.
3	Tips used for dispensing and mixing with Bravo platform were 70ul Tips (# 19133-212). Tips were freshly discharged with a Milty Zerostat3 not more than 1h before immediate use.
4	This protocol is based upon Affymetrix's protocol: QuantiGene ViewRNA HC Screening User Manual, RevD

Nature Methods: doi:10.1038/nmeth.2657

5 Few remaining manual steps could also be fully automated.

6 PreHyb plates can be refilled during the assay. This refilling is not indicated.
Prepare two plates of PreHyb to reduce effect of unexpected problems.

7 Prepare protocols for machines before starting the assay (and double check processed volumes manually).

8 For fixation, fix all wells of the plate (not only the actually used wells)

Abbreviations

RT	Room Temperature
PBS	Phosphate buffer saline
PFA	Paraformaldehyde
SuccEst	Alexa Fluor® 647 carboxylic acid, succinimidyl ester (Invitrogen)

Step	Hour	Min	Plate	Name	Description	Check	Comments
1			**N/A**	Reagents	Warm Reagents: PreHyb,PS_Diluent, Amp_Diluent, LP_Diluent,PBS (for PreHyb plates only) for 30min 40C		
2			**N/A**	Reagents	Prepare 8% PFA ad leave at RT.		
3			**N/A**	Reagents	Prepare Wash Buffer ad leave at RT.		
4			**N/A**	Reagents	Prepare Probe Sets leave at RT.		
5			**N/A**	Reagents	Make sure PreHyb is at 40C. Prepare 2 plates and store at 40C.		
6			**N/A**	Reagents	Prepare Pre Amp leave at RT.		
7			**N/A**	Reagents	Prepare Amp leave at RT.		
8			**N/A**	Reagents	Prepare Label Probe leave at RT.		
9			**N/A**	Equipment	Start **Biotek EL406** and test functionality, 1ul Cassette		
10			**N/A**	Equipment	Start **Bravo** (and **initialize Liconic)**		
11			**N/A**	Break	Have Breakfast		
12			**N/A**	Reagents	Prepare Mitotracker in Serum Free Medium. Use within 1min.		OPTIONAL
13			**All**	Mitotracker	Aspirate medium, dispense **Mitotracker** (15ul), Incubate@37C for 45 min.		OPTIONAL
14			**All**	Cell Fixation	Wash 2x with PBS (80ul), fix plate with **8% PFA** (15ul) for 30 min at RT.		
15			**All**	Cell Fixation	Aspirate PFA and wash 3x with PBS.		
16			**N/A**	Equipment	Remove **First and Last tubes** of Biotek EL406 1ul cassette.		
17			**1**	Cell Perm.	Aspirate PBS, dispense **Detergent Solution** (15ul) and incubate for 3 min at RT.		
18			**1**	Cell Perm.	Wash 2x with PBS.		
19			**1**	Protease	Aspirate PBS, dispense **Working Protease** (15ul), incubate 10 min at RT with lid closed.		
20			**1**	Protease	Wash 5x with PBS.		
21			**1**	Protease	Aspirate PBS, dispense **Protease Stop Buffer** (15ul).		
22			**1**	Protease	Mix by pipetting up and down twice		
23			**1**	Protease	Aspirate **Protease Stop Buffer**, dispense Protease Stop Buffer (15ul).		
24	**0**	**0**	**1**	Probe Set	Dis pence **Working Probe Set** with Bravo (15ul). Place Plate in incubator		
25	**0**	**10**	**2**	Cell Perm.	Aspirate PBS, dispense **Detergent Solution** (15ul) and incubate for 3 min at RT.		

Nature Methods: doi:10.1038/nmeth.2657

26			**2**	Cell Perm.	Wash 2x with PBS.		
27			**2**	Protease	Aspirate PBS, dispense **Working Protease** (15ul), incubate 10 min at RT with lid closed.		
28			**2**	Protease	Wash 5x with PBS.		
29			**2**	Protease	Aspirate PBS, dispense **Protease Stop Buffer** (15ul).		
30			**2**	Protease	Mix by pipetting up and down twice		
31			**2**	Protease	Aspirate **Protease Stop Buffer**, dispense Protease Stop Buffer (15ul).		
32			**2**	Probe Set	Dispense **Working Probe Set** with Bravo (15ul). Place Plate in incubator		
33	**0**	**50**	**3**	Cell Perm.	Aspirate PBS, dispense **Detergent Solution** (15ul) and incubate for 3 min at RT.		
34			**3**	Cell Perm.	Wash 2x with PBS.		
35			**3**	Protease	Aspirate PBS, dispense **Working Protease** (15ul), incubate 10 min at RT with lid closed.		
36			**3**	Protease	Wash 5x with PBS.		
37			**3**	Protease	Aspirate PBS, dispense **Protease Stop Buffer** (15ul).		
38			**3**	Protease	Mix by pipetting up and down twice		
39			**3**	Protease	Aspirate **Protease Stop Buffer**, dispense Protease Stop Buffer (15ul).		
40			**3**	Probe Set	Dispense **Working Probe Set** with Bravo (15ul). Place Plate in incubator		
41	**1**	**20**	**N/A**	Equipment	Change Cassette of BioTek to 5ul Cassette		
42	**1**	**25**		Break	Lunch		
43	**2**	**50**	**1**	Probe Set	Aspirate **Working Probe Set**, do 3x Wash Buffer at RT followed by 30 sec incubation.		
44			**1**	Pre Amp	Aspirate Wash Buffer, dispense **PreHyb Buffer** (15ul) with Bravo by quadrant, pipette up and down twice, aspirate PreHyb		
45			**1**	Pre Amp	Dispense **Working Pre Amp** (15ul) with Bravo by quadrant. Place plate in the incubator.		
46	**3**	**40**	**2**	Probe Set	Aspirate **Working Probe Set**, do 3x Wash Buffer at RT followed by 30 sec incubation.		
47			**2**	Pre Amp	Aspirate Wash Buffer, dispense **PreHyb Buffer** (15ul) with Bravo by quadrant, pipette up and down twice, aspirate PreHyb		
48			**2**	Pre Amp	Dispense **Working Pre Amp** (15ul) with Bravo by quadrant. Place plate in the incubator.		
49	**4**	**0**	**1**	Pre Amp	Aspirate **Working Pre Amp**, do 3x Wash Buffer at 40C followed by 30 sec incubation.		
50			**1**	Amp	Aspirate Wash Buffer, dispense **PreHyb Buffer** (15ul) with Bravo by quadrant, pipette up and down twice, aspirate PreHyb		
51			**1**	Amp	Dispense **Working Amp** (15ul) with Bravo by quadrant.		
52	**4**	**30**	**3**	Probe Set	Aspirate **Working Probe Set**, do 3x Wash Buffer at RT followed by 30 sec incubation.		

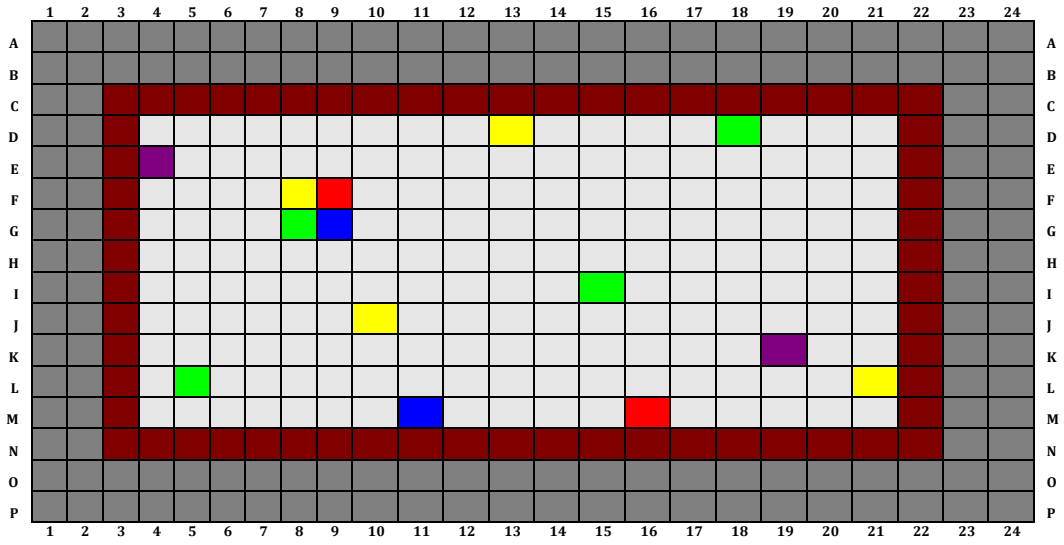
Nature Methods: doi:10.1038/nmeth.2657

53			**3**	Pre Amp	Aspirate Wash Buffer, dispense **PreHyb Buffer** (15ul) with Bravo by quadrant, pipette up and down twice, aspirate PreHyb		
54			**3**	Pre Amp	Dispense **Working Pre Amp** (15ul) with Bravo by quadrant. Place plate in the incubator.		
55	**4**	**50**	**2**	Pre Amp	Aspirate **Working Pre Amp**, do 3x Wash Buffer at 40C followed by 30 sec incubation.		
56			**2**	Amp	Aspirate Wash Buffer, dispense **PreHyb Buffer** (15ul) with Bravo by quadrant, pipette up and down twice, aspirate PreHyb		
57			**2**	Amp	Dispense **Working Amp** (15ul) with Bravo by quadrant.		
58	**5**	**10**	**1**	Amp	Aspirate **Working Amp**, do 3x Wash Buffer at RT followed by 30 sec incubation.		
59			**1**	LP	Aspirate Wash Buffer, dispense **PreHyb Buffer** (15ul) with Bravo by quadrant, pipette up and down twice, aspirate PreHyb		
60			**1**	LP	Dispense **Working Label Probe** (15ul) with Bravo by quadrant.		
61	**5**	**40**	**3**	Pre Amp	Aspirate **Working Pre Amp**, do 3x Wash Buffer at 40C followed by 30 sec incubation.		
62			**3**	Amp	Aspirate Wash Buffer, dispense **PreHyb Buffer** (15ul) with Bravo by quadrant, pipette up and down twice, aspirate PreHyb		
63			**3**	Amp	Dispense **Working Amp** (15ul) with Bravo by quadrant.		
64	**6**	**0**	**2**	Amp	Aspirate **Working Amp**, do 3x Wash Buffer at RT followed by 30 sec incubation.		
65			**2**	LP	Aspirate Wash Buffer, dispense **PreHyb Buffer** (15ul) with Bravo by quadrant, pipette up and down twice, aspirate PreHyb		
66			**2**	LP	Dispense **Working Label Probe** (15ul) with Bravo by quadrant.		
67	**6**	**20**	**1**	LP	Aspirate **Working Label Probe**, do 3x Wash Buffer at RT followed by 30 sec incubation.		
68			**1**	LP	Wash 3x with PBS.		
69	**6**	**50**	**3**	Amp	Aspirate **Working Amp**, do 3x Wash Buffer at RT followed by 30 sec incubation.		
70			**3**	LP	Aspirate Wash Buffer, dispense **PreHyb Buffer** (15ul) with Bravo by quadrant, pipette up and down twice, aspirate PreHyb		
71			**3**	LP	Dispense **Working Label Probe** (15ul) with Bravo by quadrant.		
72	**7**	**10**	**2**	LP	Aspirate **Working Label Probe**, do 3x Wash Buffer at RT followed by 30 sec incubation.		
73			**2**	LP	Wash 3x with PBS.		
74			**N/A**	Reagents	Prepare DAPI.		
75	**7**	**50**	**3**	LP	Aspirate **Working Label Probe**, do 3x Wash Buffer at RT followed by 30 sec incubation.		

Nature Methods: doi:10.1038/nmeth.2657

76			**3**	LP	Wash 3x with PBS.		
77	**8**	**0**	**All**	DAPI	Aspirate PBS, dispense **Working DAPI** reagent (70ul), incubate at RT for 10 min		
78			**All**	DAPI	Wash 3x with PBS.		
79	**8**	**20**	**N/A**	Reagents	Prepare SuccEst.		
80	**8**	**30**	**All**	SuccEst	Aspirate PBS, dispense **Working SuccEst** reagent (70ul), incubate at RT for 5 min		
81			**All**	SuccEst	Wash 3x with PBS.		
82	**8**	**40**	**All**	Cover	Cover all plates with sticky metal foil		
83	**8**	**50**	**1**	Microscope	Image plates (do not make new settings if tired)		

High-throughput in situ RNA hybridization: Plate layout example



	Wells / Plate
No Probeset	2
Hprt1	4
Erbb2	2
ß-actin	2
dapB	4
Individual library gene	##

Well not used. Treated only with buffers.
Well not used. Treated only with PBS.

Nature Methods: doi:10.1038/nmeth.2657

Supplementary Note 1: Calculation of signal-to-noise ratios.

To obtain an accurate estimate of the signal per each individual spot including those in crowded areas we performed sub-pixel 2D Gaussian fitting with **Supplementary Equation 1**

$$f(x,y) = ae^{-\left(\frac{(x-x_0)^2}{2\sigma^2} - \frac{(y-y_0)^2}{2\sigma^2}\right)}$$

where x and y represent coordinates in the in the image, x_0 and y_0 represent the center of the spot, σ is the standard deviation of the Gaussian distribution which is assumed to be the same in both dimensions, a is the amplitude of the Gaussian and e is the Euler's number. First, 100 detected spots were randomly picked per condition and a neighborhood of 50 pixels around the spot cropped for optimization. Each pixel P of every spot was then subdivided into a 2D Euclidian grid of sub-pixels p_i such that a length of each sub-pixel was given by $\|p_i\| = \frac{\|P\|}{k}$, where $\|\cdot\|$ is a 2-norm, $k = 5$ gives a total number of 25 sub-pixels p_i per each pixel P in the cropped spot. The local background B_{loc} was subtracted and all sub-pixel intensity values normalized by dividing with the maximum intensity, or amplitude A, in the given spot so that a in s.eq.1 was set to one. Parameters for s.eq.1 fitting each spot were learned using the minimization method (COBYLA)[1] implemented as a part of the NLopt non-linear optimization package (http://ab-initio.mit.edu/nlopt) by giving a constraint (**Supplementary Equation 2**)

$$S = \sum_{i=1}^{N} |r_i - m_i|, \qquad S \to min$$

where S is the optimization score, r_i and m_i are the real and modeled normalized intensity values for sub-pixel p_i and N is the total number of sub-pixels in the cropped

Nature Methods: doi:10.1038/nmeth.2657

image.

The intensity brightness b_i of each sub-pixel p_i in each spot was then estimated by $b_i = m_i.A$. The spot brightness after local background subtraction I_{lb} was then given by the integrating b_i for each spot, i.e. $I_{lb} = \sum_{i=1}^{N} b_i$. In order to estimate the signal-to-noise ratio SNr, we first computed the intensity brightness above the image background B_{im}, namely the mean intensity of pixels outside cells, for each sub-pixel p_i given by, $c_i = b_i + B_{loc} - B_{im}$. The signal-to-noise ratio snr_i of every sub-pixel b_i was then given by $snr_i = c_i/(B_{loc} - B_{im})$.

Supplementary Note 2: Experimental setup and calculation of percentage of spots detected by two probe set types and technical noise contribution.

Probe sets of type 1 and type 6 against *KIF11* and *ERBB2* transcripts were designed so that each probe pair of each type alternated along the transcript length (**Supplementary Fig. 3a**) to avoid differences in spot detection arising from possible hybridization biases between the 3'and 5'ends of the transcripts. bDNA FISH was carried out as described in the **Supplementary Protocol** and cells imaged in our high-magnification or the high-throughput set up. Type 1 probe pairs were labeled with AlexaFluor 546 tagged label probes and type 6 probe pairs with AlexaFluor 488 tagged label probes. Spots were detected for both probe types and the percentage of type 1 spots that were also detected with the type 6 probes set calculated. For the high-throughput setup we also performed cell segmentation and spot overlap was calculated for 5,289 and 5,391 single cells for probe sets targeting *KIF11* and *ERBB2,* respectively.

Nature Methods: doi:10.1038/nmeth.2657

The contributions of technical noise and biological variability to the total single cell data variation was calculated as described by Elowitz *et al.*[2] for intrinsic, extrinsic and total noise, respectively.

Supplementary Note 3: Robust spot detection for image-based transcriptomics

Illumination correction of images

We exploited the large amount of images acquired per plate and per channel in this study to learn the illumination and signal gain differences in our field of view and specific imaging set-up. For each channel we calculated the mean intensity μ_i and the standard deviation σ_i of each pixel p_i in the field of view. We then derived an overall mean intensity $\overline{\mu}$ as well as the mean standard deviation $\overline{\sigma}$ of all pixels. To correct the illumination bias we performed per-pixel z-scoring as in **Supplementary Equation 3**

$$z_i = \frac{In_i - \mu_i}{\sigma_i}$$

where z_i is the z-scored value for pixel p_i and In_i is the original intensity value for pixel p_i in a given image. The corrected intensity value C_i for pixel p_i in an image was then calculated as shown in **Supplementary Equation 4**

$$C_i = z_i.\overline{\sigma} + \overline{\mu}$$

Rescaling of images for spot detection

Image intensities were rescaled such that the image minimum, defined as the 0.01 quantile of the intensities, would become zero and the image maximum, defined as the

Nature Methods: doi:10.1038/nmeth.2657

0.995 quantile of the intensities, would become one and all other values would scale accordingly allowing for negative values and values larger than one. To make the rescaling robust against very dim images without signal or images with unusually high intensities, plate-wide limits for the rescaling were derived from bDNA sm-FISH negative and positive controls included on each plate. A total of 98 images of each *dapB* and the *HPRT1* controls were randomly selected to serve as a reference. The lower limit of the minimal intensity was set to the 0.1 quantile of the minimal intensities of the *dapB* control. The upper limit of the minimal intensity was set to the 0.8 quantile of the maximal intensities of the *dapB* control. The lower limit of the maximal intensity was set to the 0.4 quantile of the maximal intensities of the *HPRT1* control. The upper limit of the maximal intensity was set to the 0.8 quantile of the maximal intensities of the *HPRT1* control (arrows in **Supplementary Fig. 7b**). Rescaling parameters that fell outside of the boundaries were set to the respective lower or upper limits.

Spot detection

For spot detection we performed a Laplacian-of-Gaussian (LoG) filter to the rescaled images and defined all objects above a certain threshold as spots. Note that the LoG has been successfully applied by for enhancing mRNA spots before[3]. The filter size was of five pixels, which corresponds to 5x162.5nm (812.5nm). We used one global threshold for all images of one plate after ensuring that corrected intensities of individual images were comparable. The specific threshold value was chosen from 98 random images of the *HPRT1* control. We noticed that for all plates there was a wide range of thresholds were the number of spots detected per image did not change significantly, as reported previously[3]. The threshold is chosen at the end of this plateau to increase the stringency

Nature Methods: doi:10.1038/nmeth.2657

of the spot detection compared to granular background that does not indicate mRNA particles (**Supplementary Fig. 7c**). Note this did not affect the mean number of spots detected per cell (**Supplementary Fig. 7f)**. As an additional safety measure against dim images and false-positive spots, at least one pixel within a spot required to have an intensity which would be slightly above the permitted range of the minimal intensity of an image (**Supplementary Fig. 7b**). Note that this safety measure would not generally impact the spot detection (not shown). Such a global approach circumvents the manual threshold selection for individual images described previously[3]. Also it increased the window of the signal between the expressed genes and negative controls since threshold detection algorithms without plate-wide limits would have a large number of false-positive spots that do not represent mRNA spots.

Spot bias correction

When we calculated the probability of each pixel in our field of view to be a the centroid of a detected spot we realized that there was a small bias (± 2%) towards detecting spots located at the center of our field of view. Such bias is likely to arise from the large chip size of the sCMOS camera, allowing the lens curvature to have an effect (**Supplementary Fig. 7d**). The bias was corrected by locally applying different thresholds. In order to estimate the right threshold value per pixel in our field of view we applied 20-40 thresholds close the to the reference threshold for each individual plate before the main spot detection. For each threshold the amount of spots at a given position was determined and values represented as images, which were then smoothened using a Gaussian filter of 6.5um. Local scaling factors for the global threshold were determined in such a way that the local average would correspond to

Nature Methods: doi:10.1038/nmeth.2657

the mean number of spots expected from the mean at the global reference threshold. These scaling factors were again smoothed using a Gaussian of 4.9um before applying them to detect spots in order to avoid over-fitting.

Gaussian deblending of crowded spots

While the spot detection resolved most spots, it failed to properly resolve individual spots of highly abundant RNA species, such as mRNA of *ACTB*. To increase the dynamic range, intensity-based deblending was done to separate local peaks. We achieved this using a custom-made fast implementation in MATLAB of the deblending function from the astrophysics software SourceExtractor[4]. Instead of using a logarithmic scale on intensities (as is customary in astrophysics), we used a linear scale for deblending since the RNA spots tend to have similar intensities. (**Supplementary Fig. 7e**).

Supplementary Note 4: Calculation of fraction of cells with spot counts above background and mean spot per cell correction.

The fraction of cells with spot counts above the *dapB* control s_i was calculated using sampled distributions of each gene replicate normalized by the total number of cells of a given sample. For every given expression level i, when i is at least 1 spot per cell, s_i was defined as $s_i = d_i H(d_i)$, where d_i is the difference between the observed fraction of cells g_i for a given cell sample of a gene replicate and that of its corresponding *dapB* control c_i, and $H(d_i)$ is a Heaviside step function whose value is 0 for negative d_i and 1 for positive d_i. The total fraction of cells above the negative control S_g was then given by $S_g = \sum_{i=1}^{N} s_i$, where N is the maximum number of spots in one cell observed in both

Nature Methods: doi:10.1038/nmeth.2657

replica experiments. The fraction of cell s_0 that did not show expression above *dapB*, i.e when $i = 0$, was then given by $s_o = 1 - S_g$. The corrected mean expression μ_g was then obtained by $\mu_g = \frac{\sum_{i=0}^{N} s_i M}{M}$, where M is the number of sampled cells for that gene replicate. Calculation of S_g and μ_g was bootstrapped 100 times by subsampling one third of the cells for every gene replicate and the *dapB* control cells and mean $\overline{S_g}$and $\overline{\mu_g}$ obtained for every replicate gene. The fraction above negative control expected by random was obtained for every multi-well plate in the screen by 10^4 bootstrapped comparisons of two subsampled populations of the *dapB* control. The bootstrapped distribution obtained for every replicate gene was then compared to the distributions of fractions expected by random using the Mann-Whitney-Wilcoxon test and p values obtained corrected for multiple testing using the Holm-Bonferroni method. To identify genes with fractions of cells above background, we set a conservative significance value of p=10^{-4}. Note that *dapB* control wells did not reach significance even when a less conservative significance level (p=0.05) was applied.

Supplementary Note 5: Identification of cellular patterns of mRNA localization.

Feature generation

Primary spot features were obtained by the custom "MeasureLocalizationOfSpots.m" CellProfiler[5] module (see **Fig. 2** and module help on **Supplementary Software**). Of 18 primary features calculated only 16 were used for analysis:

1. Closest distance to membrane

Nature Methods: doi:10.1038/nmeth.2657

2. Distance to cell centroid

3. Distance to nuclear centroid

4. Distance to cell outline projecting nuclear centroid

5. Mean distance to remaining spots

6. S.d of distance to remaining spots

7-12. Radius in pixels to include 5%, 10%, 15%, 25% 50% or 75% of all remaining spots in a cell. 13-16. Number of spots within 20, 40, 80 and 120 pixels from a given spot centre. Note that the feature describing the identity of the closest membrane was left out because it is more likely to describe the location of the cell in a population rather than the localization patterns of spots as cells with many neighbours are more likely to have spots close to a cell-to-cell boundary. For every cell the mean and s.d. of all above features were calculated and normalized to the square root of the area of a given cell, thus generating 32 cellular features.

Selection of data and genes for analysis

As mentioned in the main text, genes were considered for analysis if the comparison of the expression level obtained by RNA-seq and the mean number of spots per cell were consistent (including mitochondria genes) and if the mean expression level was 10 or more spots per cell (442 genes). To only take into account robust features, cells that had less that 10 spots per cell were also discarded from the analysis.

Nature Methods: doi:10.1038/nmeth.2657

Cell sampling and hierarchical clustering of single cells

To make sure that sampled cells represented the full range of the 32-dimensional feature space and accounted for potential overrepresentation in mRNA localization due to more abundant cell shapes or sizes we devised a two-step sampling procedure. First, ~$7x10^4$ cells were randomly sampled from the initial pool of cells and the pairwise Euclidean distance to a randomly sampled 20% of the $7x10^4$ cells was computed in the 32-dimensional feature space after z-score normalization. The number of neighbours of every cell was then calculated. We chose a distance for the definition of a neighbour so that most cells (99.9%) had at least one neighbour. We then grouped cells in 200 bins of similar number of neighbours and randomly sampled from each bin ~50 cells. This resulted in the sampling of ~10,000 cells per cell clustering run. Hierarchical clustering of 3,960 cell-samplings was done using a Euclidian distance space and Ward's linkage method. We chose a final number of clusters so as to maximize the adjusted rand index[6] of two different classifications of the same cells. This was done as follows: the results of the clustering of a given sample of cells were taken to be the classification *A* for that given sample set. Classification *B* was obtained by allocating to every cell in the given sampling set the cluster ID of the closest neighbouring cell from a second non-redundant sampling set. The adjusted rand index was then calculated for classifications *A* and *B*. On average ~3,960 pairwise comparisons had the maximum rand index when five clusters where assumed.

Nature Methods: doi:10.1038/nmeth.2657

Definition of main localization types, computation of cellular mRNA localization phenotypes and description of gene localization patterns

To classify clusters from different sampling sets to common localization types we computed the centroid for every obtained cluster and grouped them using hierarchical clustering into five mRNA localization types. To measure the mRNA localization phenotype of a cell we measured the distance to a number of randomly sampled cluster centroids and defined the localization type of the cell as that of its closest centroid. Such classifications were run 10,000 times for every cell and the phenotype of a cell defined as the vector of fractions of times a cell was defined as belonging to a particular localization type. Note that classification using seven centroids per iteration yielded the largest reproducibility between the two replicate experiments (**Supplementary Fig. 11g**). From the above analysis the localization pattern of a given gene can be described by the distributions of the classification fraction given by a population of cells for every localization type. We describe such distributions by computing the mean, variance, and the 3rd and 4th central moments. See https://github.com/pelkmanslab/locpatterns for example MATLAB code.

Supplementary Note 6: Preparation of multivariate transcript readouts for clustering and feature selection

Feature generation

Treatment of primary spatial features: Primary spot features were obtained by the custom "MeasureLocalizationOfSpots.m" CellProfiler module (**Fig. 2**, **Online Methods** and **Supplementary Software**). Measurements of individual RNA-spots were

Nature Methods: doi:10.1038/nmeth.2657

normalized by z-scoring against random cytoplasmic pixels. The number of sampled pixels would be the same as the number of spots, which were originally identified in the given cells. Sampling was done with full randomization without pixel replacement. Sampling of pixels was iterated 100 times. Per iteration, the measurements of a single random pixel were used for the normalization of the measurements of a given observed RNA-spot. Normalization was done by assigning observed spots the corresponding z-scored value of the 100 pixel distribution.

Generation of cellular features: For primary spatial features, cellular features were obtained by the custom "MeasureChildren.m" CellProfiler module (**Supplementary Software**). For individual parent objects ("Cells"), measurements robust to NaN (Not a Number) were derived from the normalized primary features of its children. The statistics measured were mean, median, standard deviation (s.d.) and variance as well as third, fourth, fifth and sixth central moments. For individual parent objects ("Cells"), only children ("Spot") measurements that were not NaN, were considered for calculation of the central moments. If this was not possible for a given set of primary features, the cellular feature would be set to NaN. For spot count features, additional cellular features (besides the raw spot count) were obtained by log-transforming the raw spot count and by division of the raw spot count by the area of the cell (defined as pixels within 2D segmentation of nucleus and cytoplasm) and multiplication with the median area of all cells included in either replicate assay (**Supplementary Fig. 15b**).

Nature Methods: doi:10.1038/nmeth.2657

Generation of well features: Well features were obtained using a custom Matlab script following CellProfiler analysis. Measurements robust to NaN were derived from the cellular features. In this case we computed the mean, median, and variance using Matlab's built-in NaN-robust implementations thereof. NaN-robust 3rd to 6th central moments were created by considering only cells, where the measurement was not NaN. If no such cell was present, the given well measurement would be NaN. Note that a single well would correspond to a single gene within one replicate assay. Well measurements, containing an infinite number, were substituted with NaN.

Selection of data and genes for analysis

Genes were considered for analysis if the comparison of the expression level obtained by RNA-seq and mean number of spots was consistent. This was true for genes, which were no outliers in the direct comparison of mean expression levels, and for mitochondrially encoded RNAs (13 genes). In addition, the mean number of spots per cell had to be more than 10 in both replicates (442 genes).

Starting sets

Starting sets were the initial groups of features prior to feature selection and further analysis. Individual starting sets could have overlapping features. ALL starting features included mean and variability per well features of spot count and spatial features and the mean and variability of the classification of localization patterns (**Supplementary Fig. 15b**).

Nature Methods: doi:10.1038/nmeth.2657

Data pre-processing

Data pre-processing was done separately for each starting set. Features were discarded prior to feature selection, if in at least 10% of the genes of an individual assay they were NaNs (leaving 620 of 638 features of ALL starting features, **Supplementary Fig. 14b**). Remaining NaNs (561 of 548,080 datapoints of ALL starting features) were approximated using a weighted average of the corresponding feature of the five closest genes in the Euclidean space (using MATLAB's knnimpute).

Feature selection

Genes were randomly split into two equally sized non overlapping groups. The first group served as a training set for feature selection, whereas the second group served to estimate the reproducibility after feature selection (**Supplementary Fig. 14c**). Features of the training set were Z-scored separately for each replicate assay. The fraction of highly reproducible genes was determined as described in Online Methods. Individual features were removed separately. The feature in whose absence there would be the highest fraction of highly reproducible genes was chosen. In case that the absence of multiple features would yield the same highest fraction of reproducible genes, one of them was chosen randomly. Next, the chosen feature was removed. Using the same training set of genes, features were iteratively removed until no feature was left. Afterwards, the fraction of the genes of the training set that was highly reproducible after each round of feature elimination was determined. These fractions were smoothened with a sliding average over 4 consecutive rounds of feature elimination after padding the earliest and latest values by replication. The set of features with the highest value was selected. In case that multiple sets would share the highest value, the

Nature Methods: doi:10.1038/nmeth.2657

smaller feature set was selected. The feature selection was done 60 times with different randomization events and thus distinct training sets.

Union of replicate assays for clustering and overlap with STRING 9.0

Features of individual replicate assays were normalized independently by z-scoring and then averaged over both replicate assays.

Supplementary Tables

Supplementary Table 1: Spot count per cell upon gene knock-down by RNA interference

		Wells, bDNA FISH			Single cells, bDNA FISH				Fold change bDNA FISH		qPCR	
Probes	siRNA	Mean	Stdev	Number	Mean	Stdev	Number	CV	Mean	Stdev	Mean	Stdev
TFRC	none	187.937	6.878	4	187.733	72.460	13088	0.037				
TFRC	scrambled	183.485	6.526	4	183.426	75.609	12631	0.036				
TFRC	TFRC	37.895	8.322	4	38.418	22.868	12168	0.220	0.207	0.045	0.181	0.009
MYC	none	203.013	8.026	4	203.100	86.795	10900	0.040				
MYC	scrambled	161.530	16.849	4	162.352	75.063	10393	0.104				
MYC	MYC	26.099	3.974	4	26.271	33.690	5791	0.152	0.162	0.025	0.176	0.018
ERBB2	none	23.413	0.468	4	23.403	11.087	13082	0.020				
ERBB2	scrambled	17.963	1.071	4	18.037	9.525	12230	0.060				
ERBB2	ERBB2	4.199	0.340	4	4.197	3.651	7253	0.081	0.234	0.019	0.199	0.006
dapB	none	0.484	0.053	6	0.482	1.002	21094	0.109				
none	none	0.304	0.044	6	0.303	0.992	19546	0.145				

bDNA FISH and qPCR Mean Fold Change: gene specific siRNA / scrambled
number of qPCR runs 3 all conditions
CV: Coefficient of variation

Nature Methods: doi:10.1038/nmeth.2657

Supplementary Table 2: Targeted genes in the library and their Affymetrix product ID

Affymetrix Catalog#	Gene Symbol	Entrez ID
VA1-10006	IL6	3569
VA1-10007	IL8	3576
VA1-10009	ZNF189	7743
VA1-10010	BLM	641
VA1-10011	GLI	2735
VA1-10015	PDZK1	5174
VA1-10016	TFF1	7031
VA1-10018	G6PC	2538
VA1-10022	KRT19	3880
VA1-10025	PLAUR	5329
VA1-10027	SERPINE1	5054
VA1-10029	ERCC6L	54821
VA1-10032	SIDT1	54847
VA1-10040	CDKN2A	1029
VA1-10046	Stat3	6774
VA1-10047	ATG12	9140
VA1-10048	CGB	1082
VA1-10050	ESCO2	157570
VA1-10051	HEMGN	55363
VA1-10052	CA1	759
VA1-10053	HBG2	3048
VA1-10054	HBE1	3046
VA1-10061	PVT1	5820
VA1-10069	SOD1	6647
VA1-10070	COL1A1	1277
VA1-10073	UGCG	7357
VA1-10074	LMO2	4005
VA1-10076	ETS2	2114
VA1-10078	CA2	760
VA1-10079	HLA-G	3135
VA1-10080	HBG1	3047
VA1-10085	IL13RA2	3598
VA1-10096	ABL1	25
VA1-10099	INS	3630
VA1-10113	MMP2	4313
VA1-10114	MMP1	4312
VA1-10119	GAPDH	2597
VA1-10122	IL17F	112744
VA1-10123	IL17	3605
VA1-10124	HMOX1	3162
VA1-10126	DDIT3	1649
VA1-10128	BTG2	7832
VA1-10130	LGALS3	3958
VA1-10134	fn1	2335
VA1-10148	PPIB	5479
VA1-10151	KRT7	3855
VA1-10156	PPARG	5468
VA1-10159	BMP6	654
VA1-10160	ABCA1	19
VA1-10167	HCRT	3060
VA1-10168	igfbp3	3486

Affymetrix Catalog#	Gene Symbol	Entrez ID
VA1-12073	prkcq	5588
VA1-12074	PRKCD	5580
VA1-12075	prkaa2	5563
VA1-12076	PRKAA1	5562
VA1-12077	nfe2l2	4780
VA1-12078	RPL10P15	6134
VA1-12079	MCM7	4176
VA1-12080	SMAD7	4092
VA1-12081	SMAD3	4088
VA1-12082	SMAD2	4087
VA1-12083	IDH1	3417
VA1-12084	STAM2	10254
VA1-12085	STX6	10228
VA1-12086	flot1	10211
VA1-12087	PSME3	10197
VA1-12088	Actr2	10097
VA1-12089	ACTR3	10096
VA1-12090	VAMP7	6845
VA1-12091	SOS1	6654
VA1-12092	DLL1	28514
VA1-12093	rheb	6009
VA1-12094	PTK2	5747
VA1-12095	SMAD6	4091
VA1-12096	lamp1	3916
VA1-12097	KARS	3735
VA1-12098	Akt3	10000
VA1-12099	TP73	7161
VA1-12100	Src	6714
VA1-12101	Prkce	5581
VA1-12102	RAB8A	4218
VA1-12103	smad4	4089
VA1-12104	JUND	3727
VA1-12105	HSP90AA1	3320
VA1-12106	hadh	3033
VA1-12107	cttn	2017
VA1-12108	E2F1	1869
VA1-12109	cbl	867
VA1-12110	CAPN1	823
VA1-12111	akt1	207
VA1-12112	usp8	9101
VA1-12113	PLK1	5347
VA1-12114	pfn1	5216
VA1-12115	Oxa1l	5018
VA1-12116	myo6	4646
VA1-12117	MYO1E	4643
VA1-12118	ABCC1	4363
VA1-12119	Hdac1	3065
VA1-12120	Eif4g1	1981
VA1-12121	DNM2	1785
VA1-12122	cdkn2b	1030
VA1-12123	CLOCK	9575

Nature Methods: doi:10.1038/nmeth.2657

VA1-10169	HK2	3099
VA1-10174	XBP1	7494
VA1-10179	RELN	5649
VA1-10184	BDNF	627
VA1-10185	HIF1A	3091
VA1-10188	18S	
VA1-10189	BCL2	596
VA1-10196	CYP3A4	1576
VA1-10197	Apob	338
VA1-10201	ENG	2022
VA1-10203	UBC	7316
VA1-10204	FOLH1	2346
VA1-10205	cav1	857
VA1-10206	BCL2L10	10017
VA1-10214	ELAVL1	1994
VA1-10215	ADRA1A	148
VA1-10217	SP7	121340
VA1-10219	KIF11	3832
VA1-10220	FGF21	26291
VA1-10243	AHSA1	10598
VA1-10253	GFAP	2670
VA1-10263	CXCR4	7852
VA1-10265	TNFRSF12A	51330
VA1-10266	KLF2	10365
VA1-10267	TSLP	85480
VA1-10268	CD207	50489
VA1-10269	LAMP3	27074
VA1-10270	CD80	941
VA1-10271	DD3-PCA3	50652
VA1-10272	dapB	944762
VA1-10275	CSF2	1437
VA1-10300	ACTA2	59
VA1-10301	SST	6750
VA1-10302	GLP1R	2740
VA1-10303	CALCR	799
VA1-10304	CALCA	796
VA1-10313	ITGAM	3684
VA1-10322	PCSK9	255738
VA1-10324	STK33	65975
VA1-10325	EZH2	2146
VA1-10326	MTOR	2475
VA1-10327	MYC	4609
VA1-10343	GDF2	2658
VA1-10344	ANLN	54443
VA1-10345	BRRN1	23397
VA1-10346	UBE2C	11065
VA1-10347	KIF2C	11004
VA1-10348	ACP5	54
VA1-10349	APOA1	335
VA1-10351	ACTB	60
VA1-10353	MACC1	346389
VA1-10357	TGFA	7039
VA1-10358	DTR	1839
VA1-10359	PTGS2	5743
VA1-10360	CD44	960
VA1-10361	HGF	3082
VA1-10362	SPP1	6696
VA1-10363	MET	4233

VA1-12124	CLTC	1213
VA1-12125	atg5	9474
VA1-12126	Abcg2	9429
VA1-12127	Pex16	9409
VA1-12128	zfyve9	9372
VA1-12129	etf1	2107
VA1-12130	HGS	9146
VA1-12131	RABEP1	9135
VA1-12132	Rab11a	8766
VA1-12133	SDPR	8436
VA1-12134	pabpn1	8106
VA1-12135	RAB7A	7879
VA1-12136	tgfbr1	7046
VA1-12137	TAF1	6872
VA1-12138	STX1A	6804
VA1-12139	rab5c	5878
VA1-12140	rab4a	5867
VA1-12141	PPP2R2B	5521
VA1-12142	NOTCH4	4855
VA1-12143	mcm2	4171
VA1-12144	NRG1	3084
VA1-12145	GOLGA2	2801
VA1-12146	FLOT2	2319
VA1-12147	FDFT1	2222
VA1-12148	endog	2021
VA1-12149	DVL3	1857
VA1-12150	DVL2	1856
VA1-12151	DVL1	1855
VA1-12152	DNM1	1759
VA1-12153	CREBBP	1387
VA1-12154	CREB1	1385
VA1-12155	CDH1	999
VA1-12156	Casp3	836
VA1-12157	BAX	581
VA1-12158	Bad	572
VA1-12159	arrb2	409
VA1-12160	bin1	274
VA1-12161	ACVR1B	91
VA1-12162	RAB11B	9230
VA1-12163	AIFM1	9131
VA1-12164	wars	7453
VA1-12165	rab5a	5868
VA1-12166	psmb8	5696
VA1-12167	HSPA9	3313
VA1-12168	FEN1	2237
VA1-12169	FASN	2194
VA1-12170	EIF4EBP1	1978
VA1-12171	Echs1	1892
VA1-12172	AP2S1	1175
VA1-12173	AP2M1	1173
VA1-12174	CDKN1B	1027
VA1-12175	ARRB1	408
VA1-12176	NFKB1	4790
VA1-12177	SH2D2A	9047
VA1-12178	WASL	8976
VA1-12179	btrc	8945
VA1-12180	SYNJ2	8871
VA1-12181	SYNJ1	8867

Nature Methods: doi:10.1038/nmeth.2657

VA1-10364	HCRTR2	3062
VA1-10366	VCP	7415
VA1-10367	TTK	7272
VA1-10368	NUP98	4928
VA1-10369	CCNB1	891
VA1-10373	KRT5	3852
VA1-10374	GAL	51083
VA1-10376	ADIPOR2	79602
VA1-10377	ADIPOR1	51094
VA1-10381	KLK3	354
VA1-10388	AXIN2	8313
VA1-10395	PDGFRB	5159
VA1-10396	NRP1	8829
VA1-10400	PROM1	8842
VA1-10417	RPS6	6194
VA1-10418	EEF1A1	1915
VA1-10437	ACTG2	72
VA1-10441	CDX2	1045
VA1-10449	RUNX2	860
VA1-10450	COL2A1	1280
VA1-10451	AGC1	176
VA1-10452	SOX9	6662
VA1-10453	NCOA3	8202
VA1-10456	OTUD5	55593
VA1-10478	DKK1	22943
VA1-10479	BMP2	650
VA1-10481	TNF	7124
VA1-10483	FPR2	2358
VA1-10484	Pak1	5058
VA1-10511	ZNF174	7727
VA1-10512	HEXIM1	10614
VA1-10515	IL17RB	55540
VA1-10516	IL25	64806
VA1-10519	id1	3397
VA1-10521	FGFR3	2261
VA1-10531	ADRA1D	146
VA1-10532	adra1b	147
VA1-10537	ND5	4540
VA1-10544	S100a4	6275
VA1-10545	VEGFA	7422
VA1-10581	IL23A	51561
VA1-10586	LGR6	59352
VA1-10587	GPR49	8549
VA1-10603	PMEPA1	56937
VA1-10604	OR51E2	81285
VA1-10608	NKX3-1	4824
VA1-10610	ERG	2078
VA1-10611	B2M	567
VA1-10612	SPINK1	6690
VA1-10616	Ar	367
VA1-10624	ROS1	6098
VA1-10625	ALK	238
VA1-10628	CMKOR1	57007
VA1-10629	GPR101	83550
VA1-10630	GPR123	84435
VA1-10642	GLI2	2736
VA1-10647	KRT6A	3853
VA1-10649	P2RY11	5032

VA1-12182	FBP2	8789
VA1-12183	snap23	8773
VA1-12184	TNFSF10	8743
VA1-12185	SNX4	8723
VA1-12186	BECN1	8678
VA1-12187	numb	8650
VA1-12188	Yars	8565
VA1-12189	AP3B1	8546
VA1-12190	pex3	8504
VA1-12191	dyrk3	8444
VA1-12192	Soat2	8435
VA1-12193	EEA1	8411
VA1-12194	PLA2G10	8399
VA1-12195	pla2g6	8398
VA1-12196	Pip5k1b	8395
VA1-12197	fzd1	8321
VA1-12198	stam	8027
VA1-12199	fzd5	7855
VA1-12200	xpo1	7514
VA1-12201	WNT9A	7483
VA1-12202	WNT5A	7474
VA1-12203	EZR	7430
VA1-12204	vcl	7414
VA1-12205	SUMO1	7341
VA1-12206	UBE2I	7329
VA1-12207	Tgfbr2	7048
VA1-12208	TFRC	7037
VA1-12209	STX5	6811
VA1-12210	SRPR	6734
VA1-12211	srp54	6729
VA1-12212	SRP19	6728
VA1-12213	SRP14	6727
VA1-12214	SOX2	6657
VA1-12215	soat1	6646
VA1-12216	SNAP25	6616
VA1-12217	SHC1	6464
VA1-12218	sdhD	6392
VA1-12219	SdhB	6390
VA1-12220	Sort1	6272
VA1-12221	RPA1	6117
VA1-12222	rfc1	5981
VA1-12223	RASGRF1	5923
VA1-12224	RAF1	5894
VA1-12225	RAB13	5872
VA1-12226	rab5b	5869
VA1-12227	PXN	5829
VA1-12228	pex19	5824
VA1-12229	Ptx3	5806
VA1-12230	PTPN6	5777
VA1-12231	ptpn1	5770
VA1-12232	PSMD4	5710
VA1-12233	psmb9	5698
VA1-12234	Psma6	5687
VA1-12235	MAP2K1	5604
VA1-12236	mapk8	5599
VA1-12237	MAPK1	5594
VA1-12238	Prkcz	5590
VA1-12239	PRKCI	5584

Nature Methods: doi:10.1038/nmeth.2657

VA1-10656	TRIB2	28951
VA1-10721	TMPRSS2	7113
VA1-10723	AMACR	23600
VA1-10725	PCA3	50652
VA1-10733	HOXB9	3219
VA1-10745	GHSR	2693
VA1-10746	mapt	4137
VA1-10747	GPR83	10888
VA1-10748	GPR35	2859
VA1-10749	SSTR3	6753
VA1-10750	GPER	2852
VA1-10790	LRRC31	79782
VA1-10791	FLT4	2324
VA1-10816	BCL2L11	10018
VA1-10840	IL10	3586
VA1-10846	eif4e	1977
VA1-10847	L1TD1	54596
VA1-10851	PLAU	5328
VA1-10852	FSTL1	11167
VA1-10870	COL4A1	1282
VA1-10871	SEC31B	25956
VA1-10872	PECAM1	5175
VA1-10879	ENAH	55740
VA1-10880	D4S234E	27065
VA1-10881	DEF6	50619
VA1-10882	PCDHB15	56121
VA1-10885	TRPV1	7442
VA1-10886	TRPM2	7226
VA1-10904	KRT17	3872
VA1-10938	CALCRL	10203
VA1-10939	SCN9A	6335
VA1-10943	Notch2	4853
VA1-10957	MAGEA2	4101
VA1-10961	TP73L	8626
VA1-11000	LRG1	116844
VA1-11001	EREG	2069
VA1-11010	FAM123A	219287
VA1-11011	FAM123C	205147
VA1-11020	GPC3	2719
VA1-11030	MCM6	4175
VA1-11031	ARPC2	10109
VA1-11032	CLEC3B	7123
VA1-11033	MKI67	4288
VA1-11035	CXCL12	6387
VA1-11041	CCL2	6347
VA1-11042	CCR2	729230
VA1-11043	NOS2	4843
VA1-11055	RAVER2	55225
VA1-11056	TGM4	7047
VA1-11057	SEMG2	6407
VA1-11070	HRH3	11255
VA1-11072	GAD1	2571
VA1-11073	HCRTR1	3061
VA1-11084	HSPA5	3309
VA1-11095	KRT14	3861
VA1-11096	KRT10	3858
VA1-11103	PPIA	5478
VA1-11104	OLFM4	10562

VA1-12240	PPP2CA	5515
VA1-12241	POLR2G	5436
VA1-12242	POLR2E	5434
VA1-12243	PML	5371
VA1-12244	PLD2	5338
VA1-12245	PLD1	5337
VA1-12246	PIK3CG	5294
VA1-12247	SERPINB5	5268
VA1-12248	PGF	5228
VA1-12249	pfn2	5217
VA1-12250	pfkp	5214
VA1-12251	PFKL	5211
VA1-12252	Per1	5187
VA1-12253	PDPK1	5170
VA1-12254	Pak2	5062
VA1-12255	PABPC1	26986
VA1-12256	OCLN	1.01E+08
VA1-12257	MSI1	4440
VA1-12258	mmp7	4316
VA1-12259	MDM4	4194
VA1-12260	mdm2	4193
VA1-12261	MCM3	4172
VA1-12262	ldhb	3945
VA1-12263	lamp2	3920
VA1-12264	KDR	3791
VA1-12265	Jun	3725
VA1-12266	ITGA5	3678
VA1-12267	ITGA3	3675
VA1-12268	ACO1	48
VA1-12269	IDH2	3418
VA1-12270	HSPD1	3329
VA1-12271	HMGB1	3146
VA1-12272	GTF2F1	2962
VA1-12273	GSK3B	2932
VA1-12274	GRB2	2885
VA1-12275	Gars	2617
VA1-12276	FGF2	2247
VA1-12277	EPS15	2060
VA1-12278	eno2	2026
VA1-12279	Egf	1950
VA1-12280	DCN	1634
VA1-12281	CTSD	1509
VA1-12282	CTSB	1508
VA1-12283	csnk2a2	1459
VA1-12284	CSNK2A1	1457
VA1-12285	CLTA	1211
VA1-12286	Cel	1056
VA1-12287	Cdc42	998
VA1-12288	Cdk1	983
VA1-12289	CAPN2	824
VA1-12290	ATP5B	506
VA1-12291	Rhoa	387
VA1-12292	Arf6	382
VA1-12293	Arf1	375
VA1-12294	Apex1	328
VA1-12295	ampH	273
VA1-12296	akt2	208
VA1-12297	parp1	142

Nature Methods: doi:10.1038/nmeth.2657

VA1-11109	Mmp9	4318
VA1-11113	KIF20A	10112
VA1-11115	DPY19L1	23333
VA1-11121	PIK3CA	5290
VA1-11122	SPARC	6678
VA1-11123	PCGEM1	64002
VA1-11124	HPRT1	3251
VA1-11127	TBP	6908
VA1-11131	BDKRB1	623
VA1-11133	pcnA	5111
VA1-11134	BCL6	604
VA1-11135	BRAF	673
VA1-11137	BIRC5	332
VA1-11144	MYF5	4617
VA1-11152	tp53	7157
VA1-11167	GPR142	350383
VA1-11168	GCG	2641
VA1-11180	MTNR1B	4544
VA1-11181	MTNR1A	4543
VA1-11187	F2RL1	2150
VA1-11211	RPLP0	6175
VA1-11212	GUSB	2990
VA1-11257	FEV	54738
VA1-11265	P2RX7	5027
VA1-11269	MSLN	10232
VA1-11274	GPR40	2864
VA1-11276	RALA	5898
VA1-11277	RALB	5899
VA1-11281	IFNB1	3456
VA1-11286	KRT15	3866
VA1-11301	IL4	3565
VA1-11313	MYOD1	4654
VA1-11317	MALAT1	378938
VA1-11327	ISL1	3670
VA1-11387	IFI27	3429
VA1-11388	USP18	11274
VA1-11389	STAT1	6772
VA1-11391	SERPINA1	5265
VA1-11392	C11orf82	220042
VA1-11393	KRT16	3868
VA1-11397	HNF4	3172
VA1-11414	HOTAIRM1	1.01E+08
VA1-11415	HSD11B1	3290
VA1-11416	GALR2	8811
VA1-11417	KCNQ2	3785
VA1-11418	KCNQ3	3786
VA1-11425	PTP4A3	11156
VA1-11432	CHAT	1103
VA1-11454	CD68	968
VA1-11465	FAM123B	139285
VA1-11466	TG	7038
VA1-11501	ERCC1	2067
VA1-11502	TYMS	7298
VA1-11503	ESR1	2099
VA1-11504	FOS	2353
VA1-11506	KIT	3815
VA1-11548	APP	351
VA1-11552	FLI1	2313

VA1-12298	INPP1	3628
VA1-12299	gtf2i	2969
VA1-12300	Gtf2b	2959
VA1-12301	EP300	2033
VA1-12302	CELSR2	1952
VA1-12303	DYRK1A	1859
VA1-12304	DKC1	1736
VA1-12305	DAB2	1601
VA1-12306	Ctbp1	1487
VA1-12307	CSNK2B	1460
VA1-12308	CLDN3	1365
VA1-12309	CLDN4	1364
VA1-12310	AP2B1	163
VA1-12311	CAV3	859
VA1-12312	CAV2	858
VA1-12313	casp10	843
VA1-12314	Casp9	842
VA1-12315	CASP8	841
VA1-12316	CASP6	839
VA1-12317	ARHGAP5	394
VA1-12318	XIAP	331
VA1-12319	BIRC3	330
VA1-12320	Acacb	32
VA1-12321	TF	7018
VA1-12322	rps23	6228
VA1-12323	RPS19P3	6223
VA1-12324	murC	347273
VA1-12325	POLR2B	5431
VA1-12326	POLR2A	5430
VA1-12327	ABCB1	5243
VA1-12328	Igf2r	3482
VA1-12329	HMGCR	3156
VA1-12330	FOLR1	2348
VA1-12331	fgf1	2246
VA1-12332	CYP27B1	1594
VA1-12333	ALDH2	217
VA1-12334	ache	43
VA1-12335	TGFB1	7040
VA1-12336	FASLG	356
VA1-12337	SLC11A1	6556
VA1-12338	vwf	7450
VA1-12339	vhl	7428
VA1-12340	TSC2	7249
VA1-12341	Ldlr	3949
VA1-12342	Fbp1	2203
VA1-12343	CFTR	1080
VA1-12344	STK11	6794
VA1-12345	NOTCH3	4854
VA1-12346	ercc2	2068
VA1-12347	CDKN1A	1026
VA1-12348	TSC1	7248
VA1-12349	RB1	5925
VA1-12350	PPT1	5538
VA1-12351	PKLR	5313
VA1-12352	PGK1	5230
VA1-12353	PFKM	5213
VA1-12354	Pdha1	5160
VA1-12355	lipA	3988

Nature Methods: doi:10.1038/nmeth.2657

VA1-11553	CD274	29126
VA1-11566	CLU	1191
VA1-11584	KRT18	3875
VA1-11586	KRT8	3856
VA1-11600	Kras	3845
VA1-11601	CTNNB1	1499
VA1-11608	IL1R1	3554
VA1-11609	IL1B	3553
VA1-11610	Il1a	3552
VA1-11611	IL6R	3570
VA1-11612	P2RY14	9934
VA1-11634	ISG15	9636
VA1-11635	IFNG	3458
VA1-11638	IL29	282618
VA1-11654	XIST	7503
VA1-11656	VPS13A	23230
VA1-11697	TTBK1	84630
VA1-11698	TTBK2	146057
VA1-11699	GABRB3	2562
VA1-11700	GABRB2	2561
VA1-11701	GABRB1	2560
VA1-11702	GABRA2	2555
VA1-11703	GABRA1	2554
VA1-11704	NOS1	4842
VA1-11705	NOS3	4846
VA1-11709	FABP1	2168
VA1-11714	ERBB2	2064
VA1-11715	ERBB4	2066
VA1-11716	ERBB3	2065
VA1-11724	PSAT1	29968
VA1-11727	COPZ2	51226
VA1-11728	COPG2	26958
VA1-11736	egfr	1956
VA1-11737	KRT23	25984
VA1-11738	MITF	4286
VA1-11740	pten	5728
VA1-11742	TERT	7015
VA1-11753	ETV4	2118
VA1-11754	ETV5	2119
VA1-11755	PTPRC	5788
VA1-11756	CD34	947
VA1-11757	KCNK18	338567
VA1-11760	pou5f1	5460
VA1-11763	TRPA1	8989
VA1-11764	GABRA3	2556
VA1-11790	ETV1	2115
VA1-11808	ASPM	259266
VA1-11836	FZD10	11211
VA1-11837	MPO	4353
VA1-11840	CX3CR1	1524
VA1-11870	IGHG4	3503
VA1-11872	SSTR2	6752
VA1-11879	TUG1	55000
VA1-11891	rac1	5879
VA1-11893	HPSE	10855
VA1-11894	RBMY1A1	5940
VA1-11905	JAK2	3717
VA1-11906	SNORD3A	780851

VA1-12356	LCAT	3931
VA1-12357	JAK3	3718
VA1-12358	ITGB3	3690
VA1-12359	INSR	3643
VA1-12360	ICAM1	3383
VA1-12361	hadhb	3032
VA1-12362	HADHA	3030
VA1-12363	Gsn	2934
VA1-12364	NR3C1	2908
VA1-12365	GPI	2821
VA1-12366	Gla	2717
VA1-12367	GBA	2629
VA1-12368	FH	2271
VA1-12369	ercc3	2071
VA1-12370	taz	6901
VA1-12371	dld	1738
VA1-12372	CD36	948
VA1-12373	C3	718
VA1-12374	Brca2	675
VA1-12375	ATM	472
VA1-12376	Fas	355
VA1-12377	APC	324
VA1-12378	Adrb3	155
VA1-12379	ADRB2	154
VA1-12380	Psen1	5663
VA1-12381	ND3	4537
VA1-12382	ND2	4536
VA1-12383	ND1	4535
VA1-12384	CYTB	4519
VA1-12385	COX3	4514
VA1-12386	ATP6	4508
VA1-12387	COX2	4513
VA1-12388	COX1	4512
VA1-12389	ND6	4541
VA1-12390	ND4L	4539
VA1-12391	ND4	4538
VA1-12392	ATP8	4509
VA1-12393	CD209	30835
VA1-12394	Ran	5901
VA1-12395	LdhA	3939
VA1-12396	SEPP1	6414
VA1-12397	PIP5K1A	8394
VA1-12398	OGG1	4968
VA1-12399	M6PR	4074
VA1-12400	rpsA	3921
VA1-12401	FYN	2534
VA1-12402	eno1	2023
VA1-12403	Pfn3	345456
VA1-12404	rps3	6188
VA1-12405	trnL1	4567
VA1-12406	PKP4	8502
VA1-12407	TJP1	7082
VA1-12408	FOXO1A	2308
VA1-12410	HOPX	84525
VA1-12416	DLL4	54567
VA1-12423	HOXB13	10481
VA1-12424	GPR44	11251
VA1-12425	HOXB7	3217

Nature Methods: doi:10.1038/nmeth.2657

VA1-11907	RNU2-1	6066
VA1-11908	ASCL1	429
VA1-11909	SCN3A	6328
VA1-11910	SCN10A	6336
VA1-11911	NTRK2	4915
VA1-11912	NTRK3	4916
VA1-11913	NTRK1	4914
VA1-11922	PCSK5	5125
VA1-11928	NGEF	25791
VA1-11931	PITX3	5309
VA1-11932	NR4A2	4929
VA1-11944	ACCN2	41
VA1-11945	CALCB	797
VA1-11946	ADRA2C	152
VA1-11947	ADRA2B	151
VA1-11948	ADRA2A	150
VA1-11949	ACCN3	9311
VA1-11961	S100B	6285
VA1-11963	AQP4	361
VA1-11964	WNT2	7472
VA1-11965	prickle2	166336
VA1-11966	ACACA	31
VA1-11967	Pik3r1	5295
VA1-11968	CHMP4B	128866
VA1-11969	PRICKLE1	144165
VA1-11970	CHMP7	91782
VA1-11971	PRKCDBP	112464
VA1-11972	Ehd4	30844
VA1-11973	VANGL1	81839
VA1-11974	FCHO2	115548
VA1-11975	TLR4	7099
VA1-11976	sp1	6667
VA1-11977	Itgb1	3688
VA1-11978	CCND1	595
VA1-11979	Fam125b	89853
VA1-11980	Pard6b	84612
VA1-11981	vps25	84313
VA1-11982	Sh3kbp1	30011
VA1-11983	Tcf7l2	6934
VA1-11984	Map2k2	5605
VA1-11985	rufy1	80230
VA1-11986	RAB11FIP1	80223
VA1-11987	NANOG	79923
VA1-11988	arhgap10	79658
VA1-11989	CHMP6	79643
VA1-11990	pla2g4a	5321
VA1-11991	CTBP2	1488
VA1-11992	MICAL1	64780
VA1-11993	SMURF2	64750
VA1-11994	ELOVL4	6785
VA1-11995	ZFYVE20	64145
VA1-11996	relA	5970
VA1-11997	CLDN1	9076
VA1-11998	VPS18	57617
VA1-11999	RPTOR	57521
VA1-12000	RAB22A	57403
VA1-12001	SMURF1	57154
VA1-12002	CHMP1B	57132

VA1-12427	IGLC	
VA1-12428	DNTT	1791
VA1-12429	CTCFL	140690
VA1-12430	CTCF	10664
VA1-12435	PAF1	54623
VA1-12445	TCN1	6947
VA1-12450	IGF2	3481
VA1-12451	EPCAM	4072
VA1-12452	CREB3L1	90993
VA1-12482	THY1	7070
VA1-12522	JAG1	182
VA1-12523	PDPN	10630
VA1-12537	CLEC4C	170482
VA1-12538	TLR7	51284
VA1-12539	PTPRN	5798
VA1-12540	PME-1	51400
VA1-12542	CCR7	1236
VA1-12546	IL15	3600
VA1-12550	IL15RA	3601
VA1-12551	IL2RB	3560
VA1-12569	CXCL17	284340
VA1-12570	WT1	7490
VA1-12580	MGMT	4255
VA1-12582	FXN	2395
VA6-10620	GLS	2744
VA6-10768	RORC	6097
VA6-10774	HOXD4	3233
VA6-10914	PRR11	55771
VA6-11198	ID2	3398
VA6-11199	ZNF691	51058
VA6-11201	TCF4	6925
VA6-11203	MXI1	4601
VA6-11204	KLF7	8609
VA6-11205	KLF3	51274
VA6-11244	DUSP6	1848
VA6-11245	DUSP5	1847
VA6-11248	CDH2	1000
VA6-11271	ATOH1	474
VA6-11305	SLC5A5	6528
VA6-11322	WNT11	7481
VA6-11410	NES	10763
VA6-11424	CD79B	974
VA6-11427	QARS	5859
VA6-11428	NKX2-5	1482
VA6-11429	GATA4	2626
VA6-11434	EML4	27436
VA6-11439	WNT10A	80326
VA6-11453	PDCD1LG2	80380
VA6-11535	ENY2	56943
VA6-11536	TMEM141	85014
VA6-11537	STK3	6788
VA6-11538	HRSP12	10247
VA6-11539	HNRNPR	10236
VA6-11540	HOXA1	3198
VA6-11541	RECQL	5965
VA6-11542	MX2	4600
VA6-11543	CDC20	991
VA6-11544	PRIM2	5558

Nature Methods: doi:10.1038/nmeth.2657

VA1-12003	CLDN2	9075
VA1-12004	Vangl2	57216
VA1-12005	Sh3glb2	56904
VA1-12006	Pard3	56288
VA1-12007	ddit4	54541
VA1-12008	Vps35	55737
VA1-12009	EPN3	55040
VA1-12010	OXSM	54995
VA1-12011	notch1	4851
VA1-12012	myh2	4620
VA1-12013	PARD6A	50855
VA1-12014	DLL3	10683
VA1-12015	polA1	5422
VA1-12016	RAB8B	51762
VA1-12017	Chmp5	51510
VA1-12018	SNX9	51429
VA1-12019	VPS28	51160
VA1-12020	RAB4B	53916
VA1-12021	HSD17B12	51144
VA1-12022	VPS36	51028
VA1-12023	SBDS	51119
VA1-12024	ldlrap1	26119
VA1-12025	DNM3	26052
VA1-12026	KANK2	25959
VA1-12027	RAB11FIP5	26056
VA1-12028	FCHO1	23149
VA1-12029	arhgap26	23092
VA1-12030	EPN2	22905
VA1-12031	Rab11fip2	22841
VA1-12032	SNAP91	9892
VA1-12033	DNAJC6	9829
VA1-12034	ZFYVE16	9765
VA1-12035	RAB11FIP3	9727
VA1-12036	Ehd2	30846
VA1-12037	ehd3	30845
VA1-12038	ZNRD1	30834
VA1-12039	ARFGAP3	26286
VA1-12040	Mcat	27349
VA1-12041	CHMP2A	27243
VA1-12042	snx5	27131
VA1-12043	celsr1	9620
VA1-12044	vamp2	6844
VA1-12045	AP2A1	160
VA1-12046	CHMP4A	29082
VA1-12047	Chmp2b	25978
VA1-12048	epn1	29924
VA1-12049	Ap2a2	161
VA1-12050	PTRF	284119
VA1-12051	FZD4	8322
VA1-12052	BRCA1	672
VA1-12053	GABARAP	11337
VA1-12054	ACOT7	11332
VA1-12055	vps45	11311
VA1-12056	snf8	11267
VA1-12057	CHEK2	11200
VA1-12058	pemt	10400
VA1-12059	PICALM	8301
VA1-12060	TRIO	7204

VA6-11568	MLANA	2315
VA6-11569	TYR	7299
VA6-11575	NUPR1	26471
VA6-11583	UGT2B15	7366
VA6-11598	TUBB3	10381
VA6-11599	MAP2	4133
VA6-11616	RMRP	6023
VA6-11761	PAX6	5080
VA6-11766	BMF	90427
VA6-11813	DBH	1621
VA6-11880	NRON	641373
VA6-11917	TLR9	54106
VA6-11927	GPRC5C	55890
VA6-11934	SLC6A3	6531
VA6-12440	SLC16A3	9123
VA6-12444	ZBED2	79413
VA6-12528	EPM2AIP1	9852
VA6-12565	FOXP3	50943
VA8-10340	GCLM	2730
VA8-10398	SMAD1	4086
VA8-10561	NKX6-1	4825
VA8-10621	GRIK2	2898
VA8-10777	HOXD3	3232
VA8-10856	NEAT1	283131
VA8-10857	STL	7955
VA8-10858	MEG3	55384
VA8-10916	COX11	1353
VA8-10918	CENPW	387103
VA8-11009	MAPK14	1432
VA8-11053	RLBP1	6017
VA8-11138	MAPKAPK2	9261
VA8-11615	PAX8AS	654433
VA8-12460	WDR78	79819
VA8-12461	GALNT10	55568
VA8-12462	ZNF547	284306
VA8-12463	GFRA1	2674
VA8-12464	GNAS	2778
VA8-12465	BAZ1B	9031
VA8-12467	CLSTN2	64084
VA8-12468	HCN1	348980
VA8-12469	CORIN	10699
VA8-12470	TRAF6	7189
VA8-12471	DMBT1	1755
VA8-12472	TNKS	8658
VA8-12473	GML	2765
VA8-12474	IFNAR1	3454
VA8-12475	IFNA2	3440
VA10-10295	CRTC2	200186
VA10-10296	CNR1	1268
VA10-10336	HSPA1A	3303
VA10-10339	SQSTM1	8878
VA10-10770	TSIX	9383
VA10-10772	TERF2IP	54386
VA10-10775	CCNDBP1	23582
VA10-10778	PDZK1IP1	10158
VA10-10779	NCOA4	8031
VA10-10781	GYPB	2994
VA10-10783	TBXA2R	6915

Nature Methods: doi:10.1038/nmeth.2657

VA1-12061	CLTCL1	8218
VA1-12062	Cltb	1212
VA1-12063	SRP72	6731
VA1-12064	SOS2	6655
VA1-12065	rab31	11031
VA1-12066	EHD1	10938
VA1-12067	PNPLA6	10908
VA1-12068	exoc5	10640
VA1-12069	STAMBP	10617
VA1-12070	NXF1	10482
VA1-12071	XRCC1	7515
VA1-12072	Tsg101	7251

VA10-10829	BCR	613
VA10-10955	SLC17A7	57030
VA10-11037	MX1	4599
VA10-11247	CSPG4	1464

References

1.Powell, M. Direct search algorithms for optimization calculations. *Acta Numerica* (1998).

2. Elowitz, M., Levine, A., Siggia, E. & Swain, P. Stochastic gene expression in a single cell. *Science (New York, N.Y.)* **297**, 1183-6 (2002).

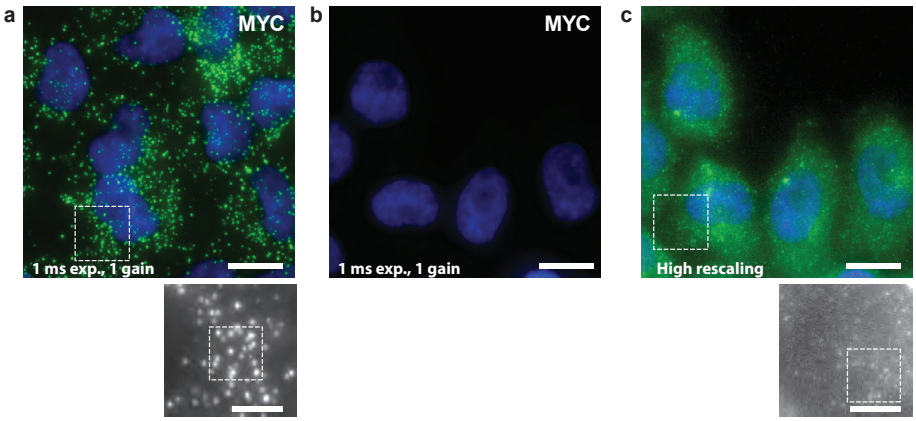
3. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature methods* **5**, 877-9 (2008).

4. Bertin, E. & Arnouts, S. SExtractor: Software for source extraction. *Astronomy and Astrophysics Supplement* (1996).

5. Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology* **7**, R100 (2006).

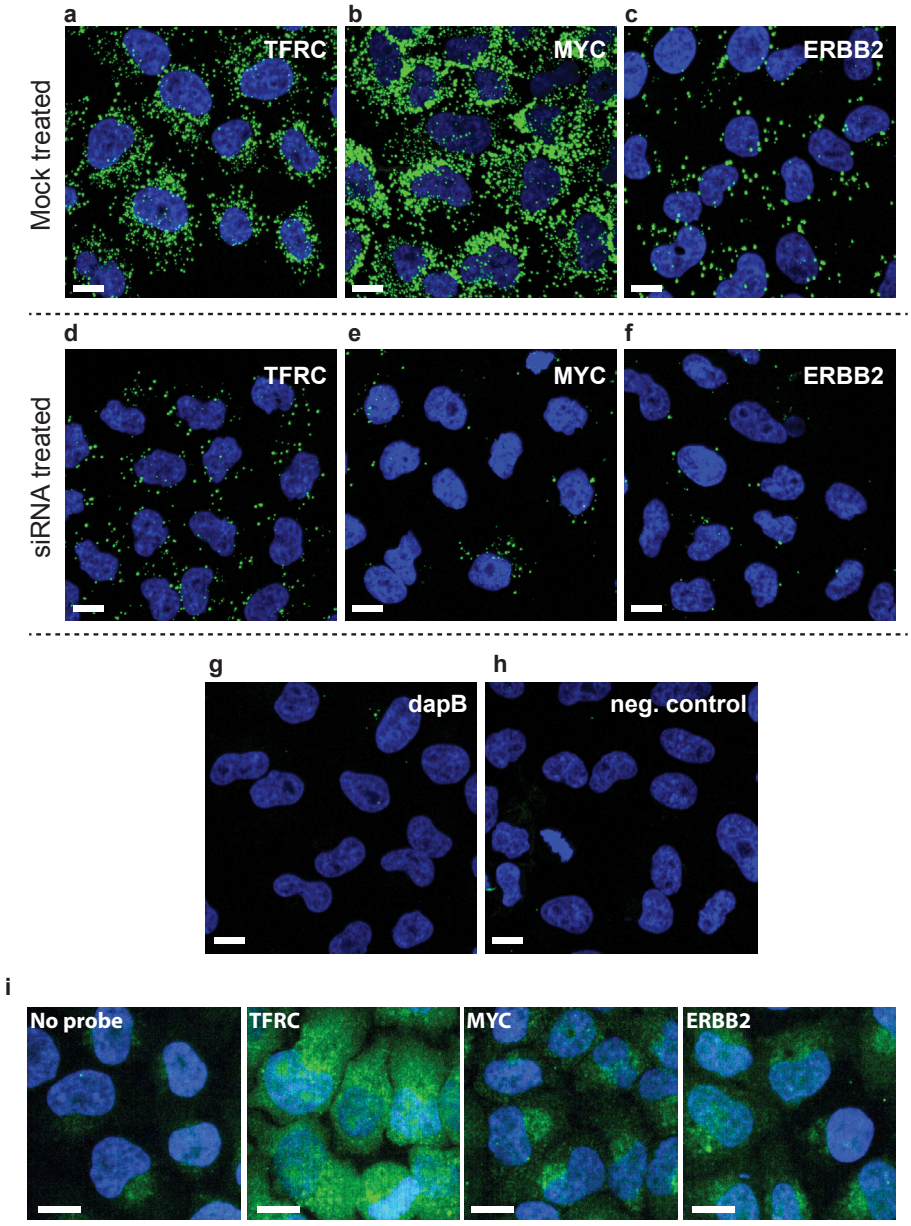
6. Martinez, W. L. & Martinez, A. R. *Computational Statistics Handbook with MATLAB, Third Edition.* (Chapman & Hall/CRC, London, UK, 2008)

Nature Methods: doi:10.1038/nmeth.2657



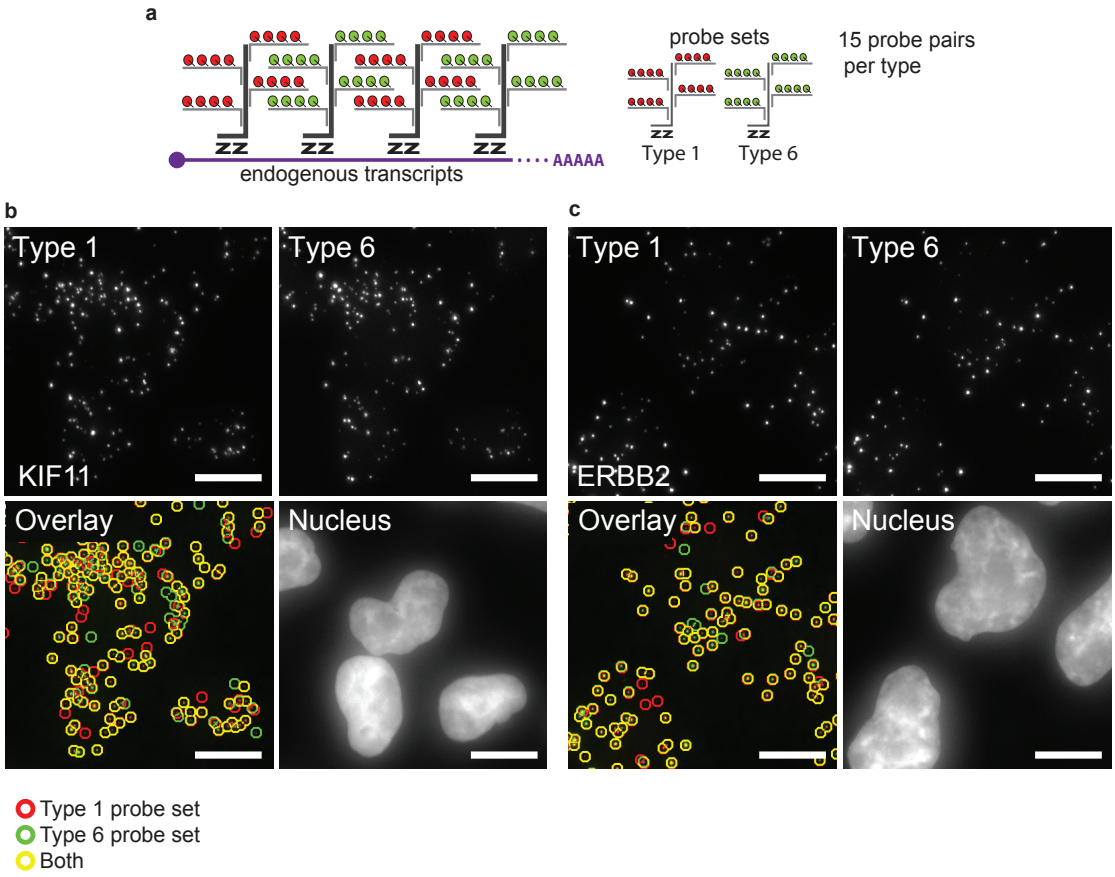
Supplementary Figure 1 | At similar imaging settings bDNA sm-FISH gives brighter signal than o-nuc sm-FISH. (**a**) Endogenous *MYC* transcript detected using bDNA sm-FISH (green) acquired with 100x magnification oil-immersion objective with 1ms exposure time and the camera gain set to one. The area marked in the lower panel corresponds to **Fig. 1e**. DAPI (cell nucleus) is blue. Scale bar: 13μm for main image and 5μm for insert. (**b**) As in **a,** but for o-nuc sm-FISH. Note that **a** and **b** were rescaled with the same parameters. (**c**) As in **b** but with a higher rescaling of the image. The area marked in the lower panel corresponds to **Fig. 1f**.

Nature Methods: doi:10.1038/nmeth.2657



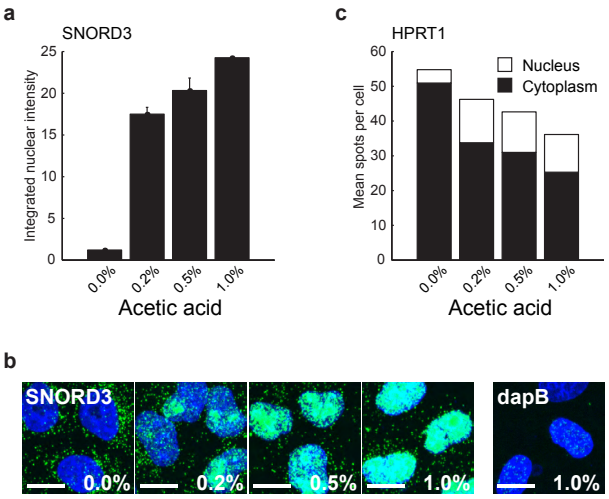
Supplementary Figure 2 | Signal from bDNA sm-FISH can be visualized with a high-throughput automated microscope and is specific for the targeted gene. HeLa cells were incubated for three days with siRNA targeting *TFRC*, *MYC* and *ERBB2*, then stained by bDNA sm-FISH (green) with probes targeting the respective genes. (**a-c**) Mock treated control cells hybridized with *TFRC*, *MYC* and *ERBB2* probes, respectively. (**d-f**) siRNA treated cells hybridized with *TFRC*, *MYC* and *ERBB2* probes, respectively. (**g,h**) Negative controls in mock treated cells using probes against the *E. coli* gene *dapB*, **g**, and a control without primary probe sets, **h**. (**i**) Acquisition of cells stained with o-nuc sm-FISH with probes against indicated transcripts and imaged with the high-throughput automated microscope (using 5 times longer exposure time than for bDNA sm-FISH). All images were acquired at 40x magnification with an automated spinning-disk microscope. DAPI (cell nucleus) is blue. Scale bars: 13μm.

Nature Methods: doi:10.1038/nmeth.2657



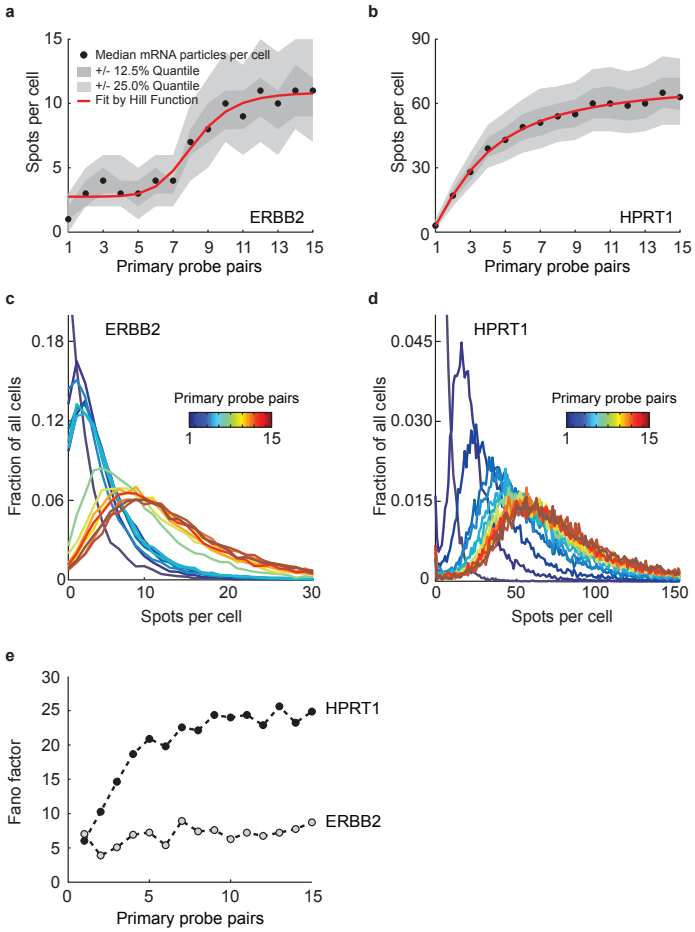
Supplementary Figure 3 | The overlap obtained by two bDNA FISH probe sets on endogenous transcripts is high. (**a**) Experimental setup with intercalating probe types along the transcript. Type 1 and 6 from Affymetrix were used. (**b**,**c**) Example images for the overlap obtained when *KIF11* and *ERBB2* endogenous transcripts where targeted using approach in **a**. Red circles: type 1 probe sets. Green circles: type 6 probe sets. Scale bar: 13μm.

Nature Methods: doi:10.1038/nmeth.2657



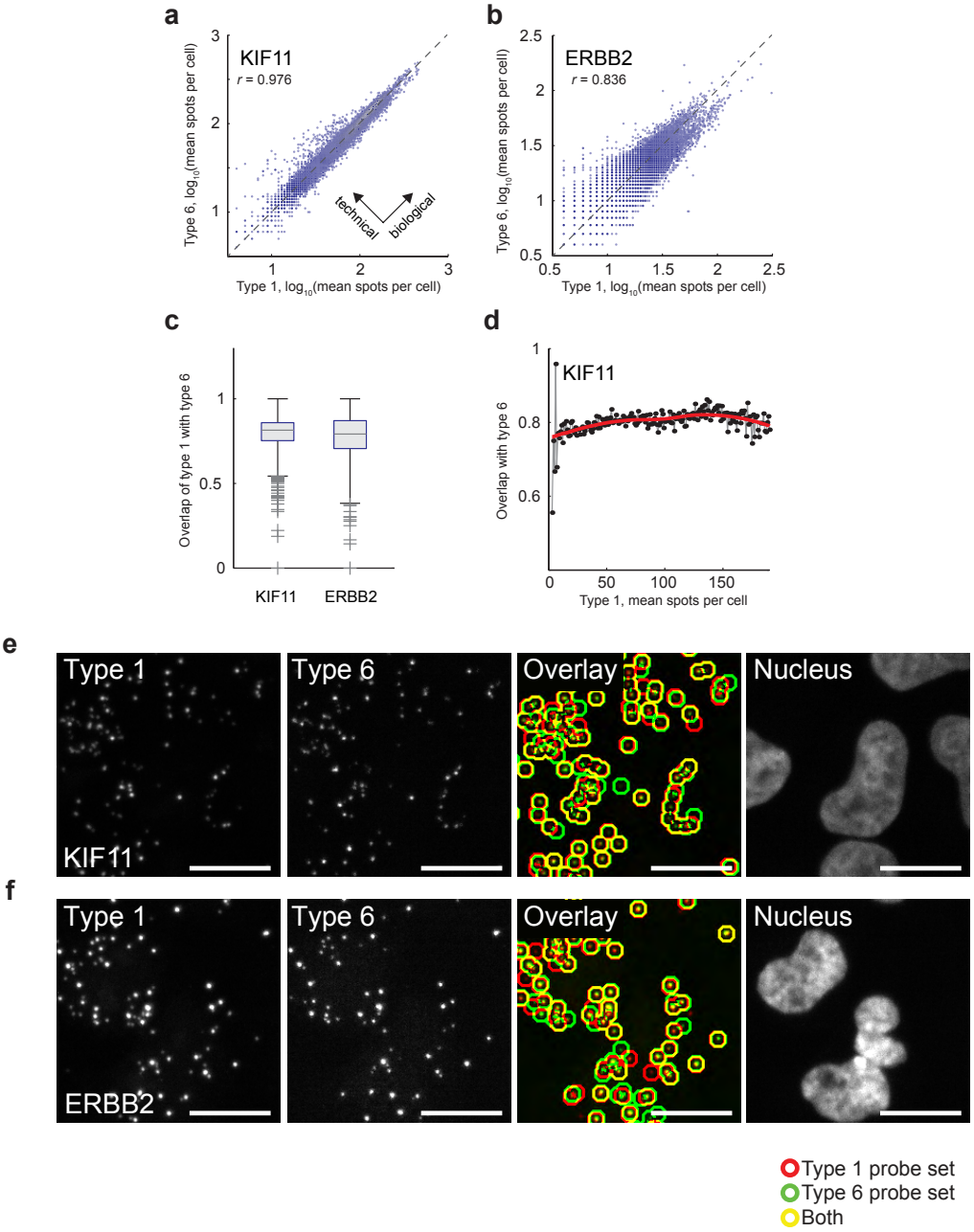
Supplementary Figure 4 | Acetic acid treatment improves nuclear accessibility but decreases staining in the cytoplasm. (**a,b**) Staining integrated intensity in the nucleus for *SNORD3* increases upon addition of acetic acid to the fixation solution (4% paraformaldehyde in PBS). Quantification of nuclear intensities is shown in **a**, while their corresponding example images are shown in **b**. bDNA sm-FISH is green, DAPI (cell nucleus) is blue. Error bars: s.d. Scale bar: 13μm. (**c**) Decrease of *HPRT1* cytoplasmic spots caused by addition of acetic acid to the fixation solution (back bars) and increase of *HPRT1* nuclear spots caused by addition of acetic acid to the fixation solution (white bars). Bar graph represents the mean of two replicates.

Nature Methods: doi:10.1038/nmeth.2657



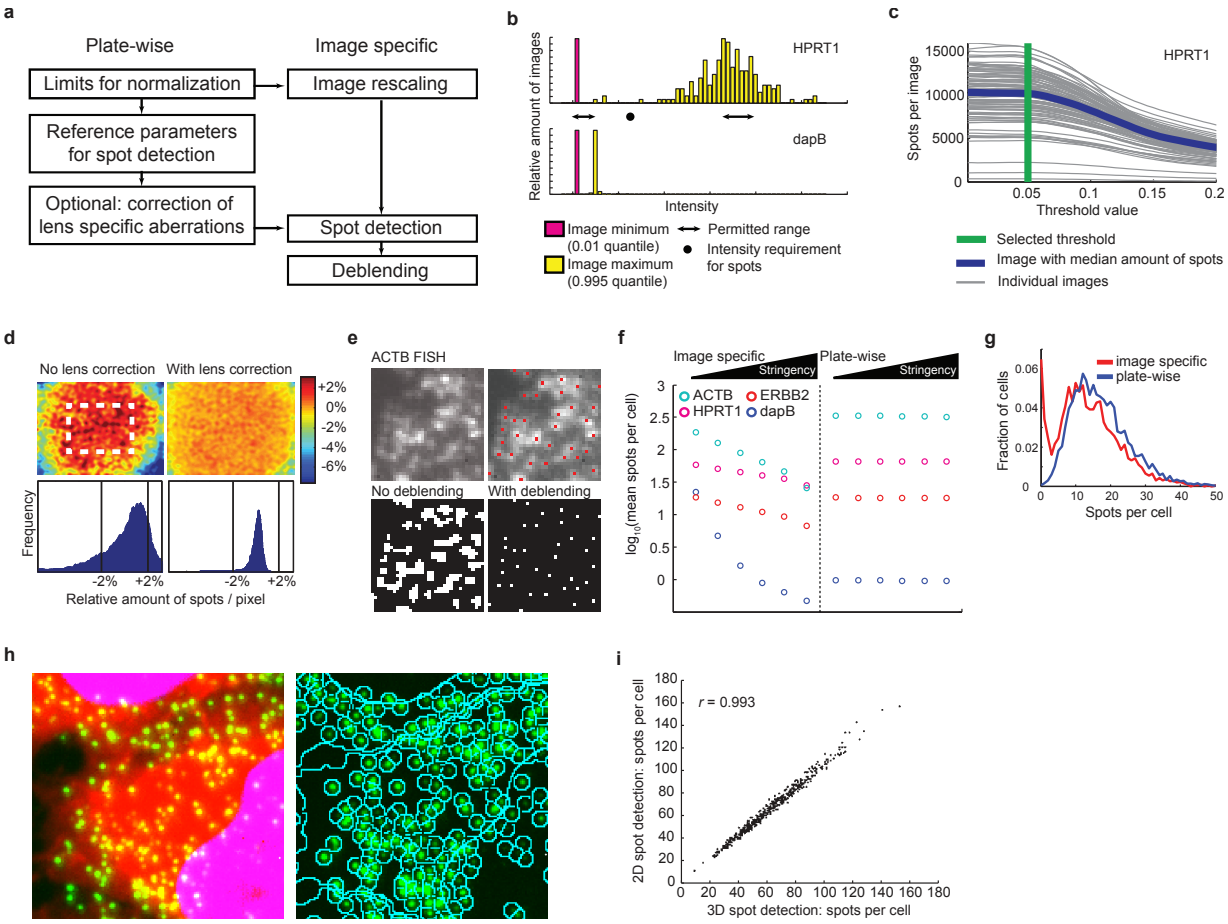
Supplementary Figure 5 | 10-15 probe pairs are required for accurate measurement of the mRNA number per cell using the bDNA FISH method. (**a**) The number of probe pairs targeting *ERBB2* was systematically increased (horizontal axis) and the number of mRNA spots per cell measured. Spot count per cell reached 80% saturation at nine probe pairs and 99% saturation at 15 probe pairs. (**b**) As in **a** but for the higher expressed transcript *HPRT1*. Spot count per cell reached 80% saturation at 10 probe pairs and 90% saturation at 15 probe pairs. (**c**) Distribution of mRNA spot counts per cell in samples stained with different numbers of probe pairs for *ERBB2*. (**d**) As in **c** but for *HPRT1*. (**e**) Fano factor (variance/mean) obtained for distributions shown in **c** and **d**.

Nature Methods: doi:10.1038/nmeth.2657



Supplementary Figure 6 | High-throughput bDNA sm-FISH is reproducible at the single cell level. (**a**,**b**) Correlation at the single cell level obtained targeting the same endogenous transcript with two different probe sets (**Supplementary Fig. 3a**) for *KIF11* and *ERBB2*, respectively. Pearson correlation value is shown, n = 10,223 cells and n = 10,524 cells for *KIF11* and *ERBB2*, respectively. The technical noise is given by the mean relative difference of spot counts per cell measured by type 1 and type 6 probe sets in a single cell, while the biological variability represents the difference of spot counts among different cells. (**c**) The fraction of detected spots by the type 1 probe set that were also detected by the type 6 probe sets in single cells, for *KIF11* and *ERBB2* transcripts, n = 5,289 and n = 5,391 cells, respectively. (**d**) The fraction of detected spots by the type 1 probe set that were also detected by the type 6 probe sets in single cells as a function of the type 1 spots per cells for *KIF11*. (**e**,**f**) Example images for the overlap obtained when *KIF11* and *ERBB2* endogenous transcripts where targeted using approach in **Supplementary Fig. 3a**. Red circles: type 1 probe sets. Green circles: type 6 probe sets. Scale bar: 13μm.

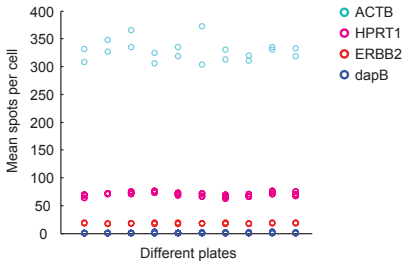
Nature Methods: doi:10.1038/nmeth.2657



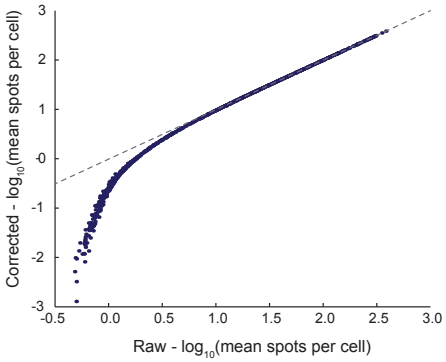
Supplementary Figure 7 | Plate-wise spot detection leads to a larger signal window between negative and positive controls and to more reliable transcript abundance distributions. (**a**) Representation of the plate-wise spot detection methodology. Limits for image rescaling, as well as reference parameters for spot detection and correction images for lens-specific aberrations are learned for each multi-well plate and applied to every image. (**b**) Distribution of image intensity minima and maxima for *HPRT1*, upper panel, and *E. coli dapB*, lower panel, are shown. Arrows indicate the permitted range for the image rescaling parameters, and the dot indicates the minimum intensity required for a pixel to be defined as a spot. (**c**) Selection of the threshold value for an example plate. The spot per image at different thresholds for a selection of images with cells hybridized with *HPRT1* probes are shown. (**d**) The bias introduced by lens aberrations is shown in the left image and histogram, the right image and histogram shows the after applying a filter to correct for position bias in the image. (**e**) Separation of spots detected in the high-expressed transcript *ACTB* using the SourceExtractor deblending algorithm. (**f**) Comparison of image specific and plate-wise spot detection. At similar levels of background spots in the *dapB* and no probes controls, plate-wise spot detection detects more spots for the three genes tested, *ACTB*, *HPRT1* and *ERBB2,* and is more robust to the stringency in the threshold. Image specific spot detection was carried out with the CellProfiler module *IdentifyPrimLog.m*. (**g**) The distribution of spots per cell was measured for *ERBB2* using image specific spot detections and plate-wise spot detections. The image-specific approach overestimates the number of cells with no mRNA spots. (**h**) Example of spot detection results in a single cell. bDNA sm-FISH is green, DAPI (cell nucleus) is blue, succinimidyl ester (cell outline) in red. (**i**) Comparison of 3D spot

Nature Methods: doi:10.1038/nmeth.2657

detection, i.e using 10 z-stacks, and 2D spot detection, using a maximum intensity projection image of the stacks. The Pearson correlation is 0.993 (n = 15,238), measurements from 500 randomly selected cells are shown.

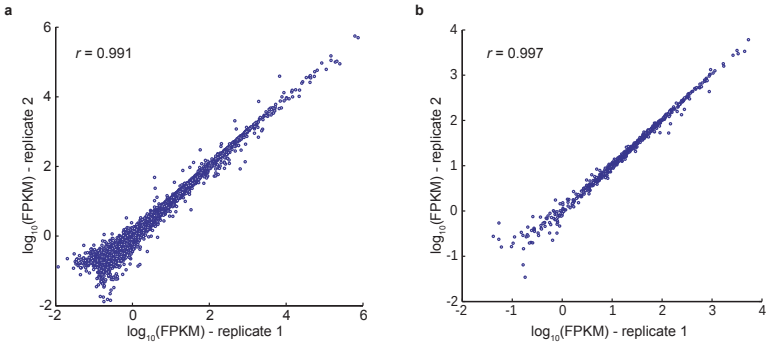


Supplementary Figure 8 | Gene expression levels of positive and negative controls are highly reproducible across plates in high-throughput bDNA sm-FISH. The logarithm of mean spots per cell in the positive controls (*ACTB*, *HPRT1* and *ERBB2*) and the negative control (*dapB*) is shown for different plates in the screen.



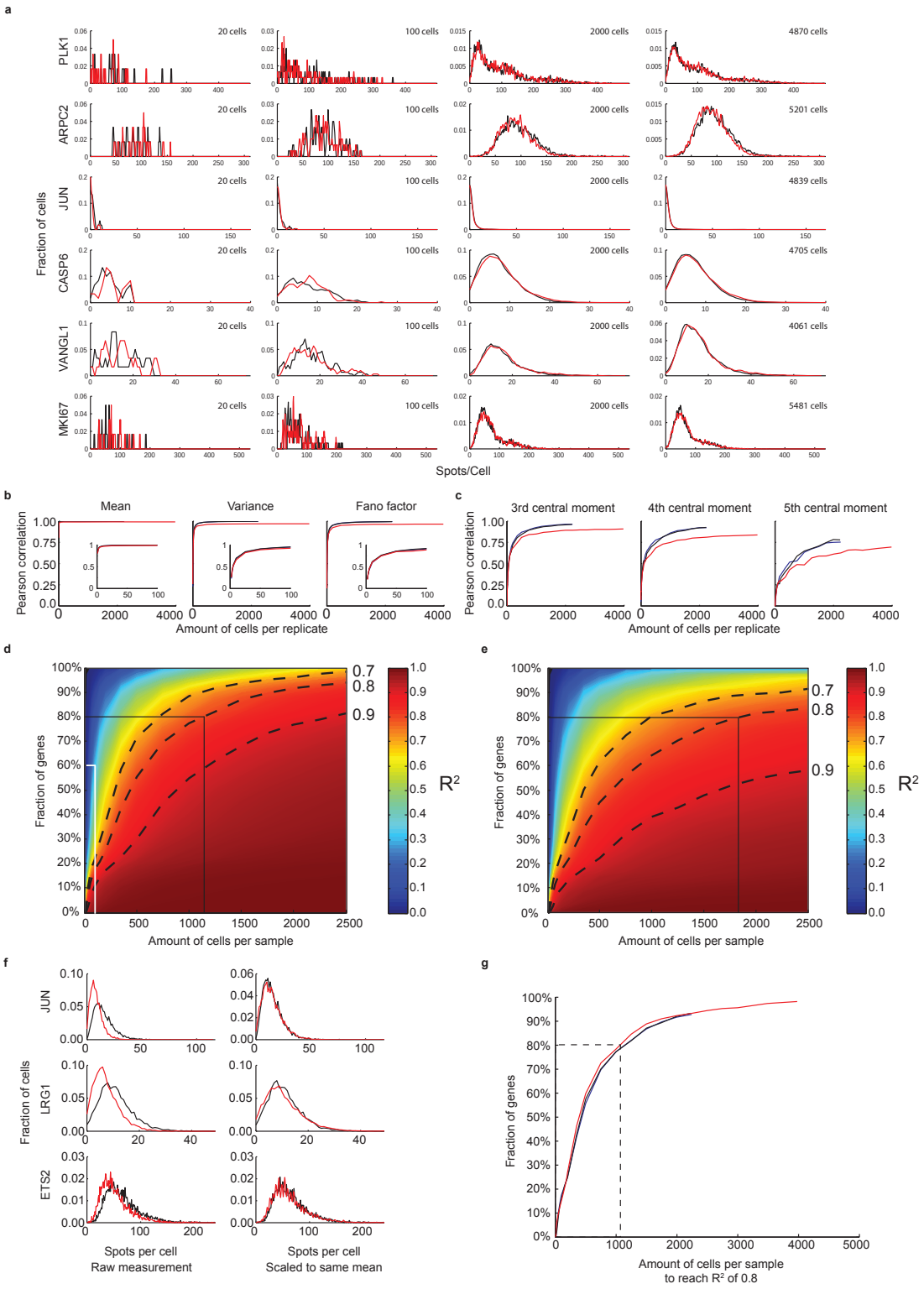
Supplementary Figure 9 | Background correction of mean spot count per cell only affects low-expressed genes. Correcting the mean spot count per cell by computing the fraction of cells above the *dapB* control is shown for the 928 genes targeted by the bDNA sm-FISH library. Broken line represent the isocline where the corrected $\log_{10}$(mean spot per cell) equals the raw $\log_{10}$(mean spots per cell) values.

Nature Methods: doi:10.1038/nmeth.2657



Supplementary Figure 10 | RNA-seq of HeLa cells is highly reproducible. (**a**) Reproducibility of RNA-seq values over the full data set. (**b**) Reproducibility of RNA-seq values over the 928 genes selected for the library. Pearson correlation values are shown.

Nature Methods: doi:10.1038/nmeth.2657

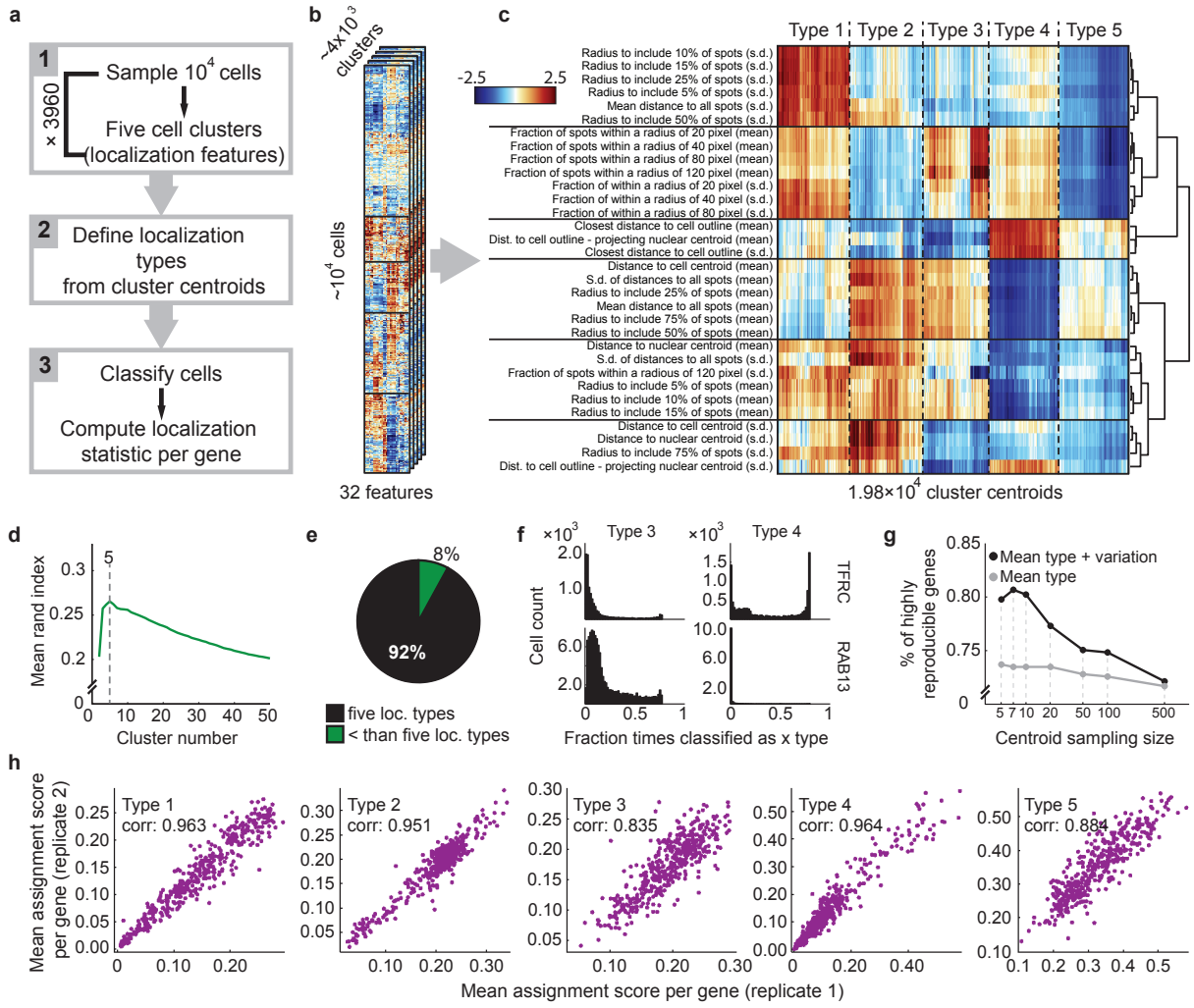


Supplementary Figure 11 | High-throughput bDNA sm-FISH reveals the minimum number of cells required for reproducible single-cell distributions of transcript abundance. (**a**) Example distributions of transcript abundance for two biological replicates (black and red lines) of different genes when sampling 20, 100, 2,000 or over 4,000 cells. (**b**) Reproducibility of the mean, variance and Fano factor (variance/mean)

Nature Methods: doi:10.1038/nmeth.2657

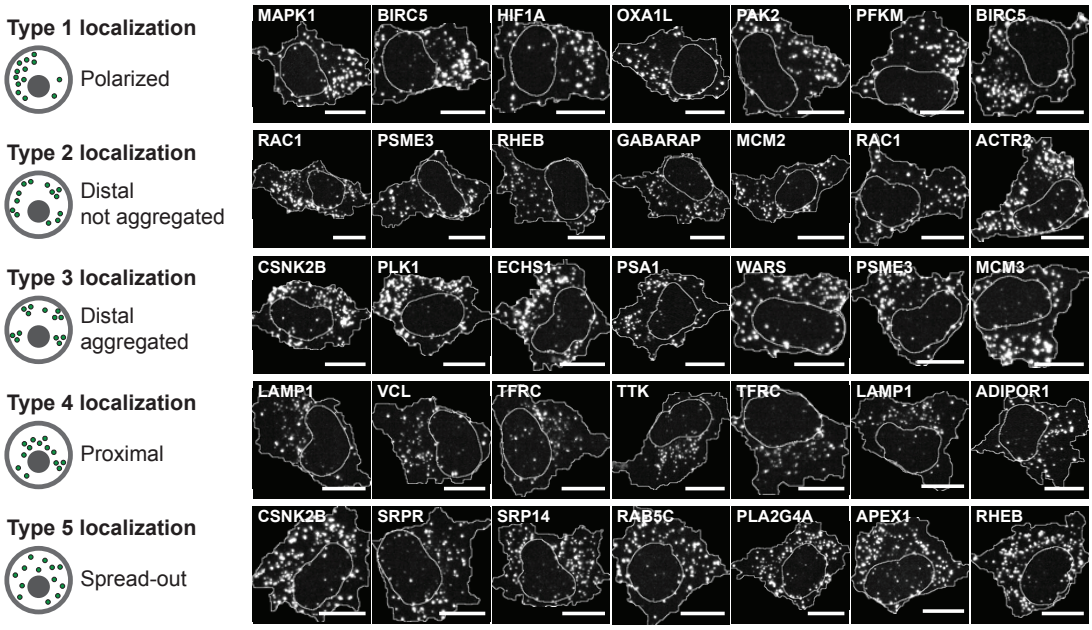
as function of the number of cell sampled comparing samples derived from the same replicates (blue and black lines) or two different replicates (red line). Lines represent the median of 100 bootstrapped samplings for each condition. (**c**) As in **b** but for the 3rd, 4th, and 5th, central moments. (**d-e**). Coefficients of determination (R^2) between single-cell spot count distributions (at a bin size of 1 spot) comparing two distributions sampled within the same biological replicate, **d**, or between the two biological replicates, **e**. The percentage of genes showing a particular R^2 as a function of the number of cells sampled is shown. Isoclines for R^2 of 0.7, 0.8 and 0.9 are shown as broken lines. Black solid lines show the amount of cells required to achieve an R^2 of 0.8 or higher for ~80% of genes. White solid line in **d** shows the R^2 achieved (<0.5) for 60% of genes if only 100 are sampled. (**f**) The distribution shape between the two replicates is conserved after correction of spot per cells measurements with the mean spots number per cell of each replicate. (**g**) The isocline corresponding to an R^2 of 0.8 is shown for samplings within the same replicate (black and blue lines) and or between the two replicates after correction for differences in the mean spot per cell (red line). Lines represent the median of 100 bootstrapped samplings for each condition.

Nature Methods: doi:10.1038/nmeth.2657



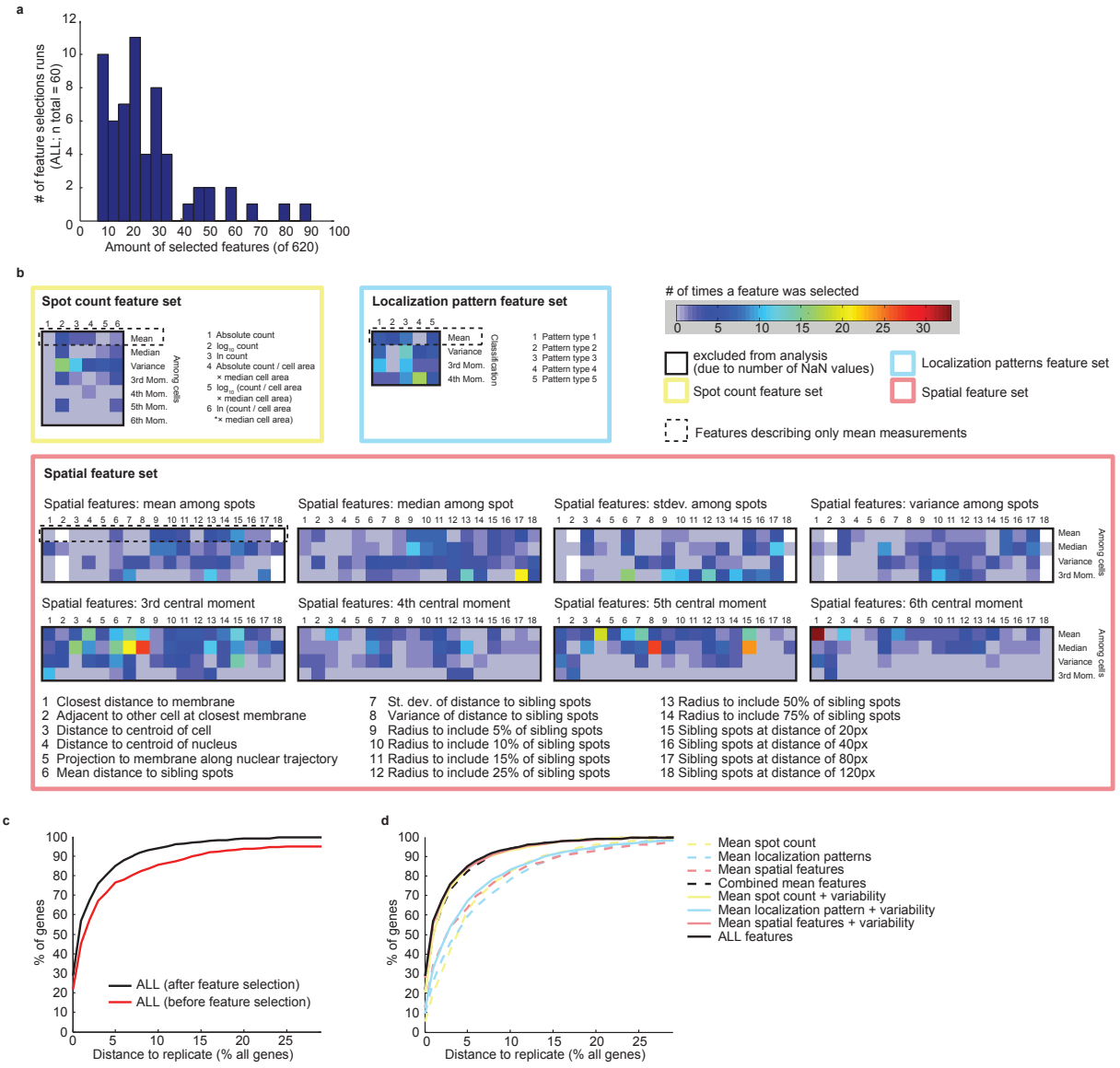
Supplementary Figure 12 | Identification of patterns of mRNA subcellular localization. (**a**) Approach taken to identify patterns of subcellular localization (**Supplementary Note 5**). (**b**) Example of hierarchical clustering result of ~10^4 sampled cells into five clusters, the sampling was repeated ~4×10^3 times with replacement. This resulted in 1.98×10^4 cluster centroids when each hierarchical clustering was divided in five clusters, see d. (**c**) Definition of localization types by clustering of centroids from **b**. (**d**) Mean adjusted rand index at different partitions of clusters in **b** for two classifications of the same sampled cells (n = 3,960). Dashed line indicates the partition number at which the maximum adjusted rand index was achieved. (**e**) Percentage of clustering runs in **b** that show cluster centroids classified as five different pattern types (black) or less than five (green). (**f**) Distributions obtained at the cell population level for the fraction of times a single cell is classified as type 2 and type 4 for two example genes, *TFRC* and *RAB13*. (**g**) The effect of the centroid sampling size when assigning cell to a given localization type on the % of genes for which their transcript patterns are classified as reproducible (where the replicate gene is within the 10% closest genes of the replicate experiment). Taking into account the variability features describing the variability of the classification distributions in **f** (variance, 3rd and 4th central moments) increases reproducibility between replicates dramatically. Dashed lines are visual guides. (**h**) Correlation plots between the mean classification score for every pattern type between replicate 1 and 2.

Nature Methods: doi:10.1038/nmeth.2657



Supplementary Figure 13 | Interpretation of types of mRNA subcellular localization patterns. Left panel shows the interpretation of the five main types of subcellular localization patterns of transcripts determined in **Supplementary Fig. 11c**. The right panel shows seven examples of cells for every pattern type. All depicted cells were classified with the respective pattern type for the majority of classification iterations, and are within the 20 closest cells to the respective centroids of the pattern types in the 32-feature space used for clustering. bDNA sm-FISH is grey. Lines indicate cell segmentation. Scale bar: 13µm.

Nature Methods: doi:10.1038/nmeth.2657

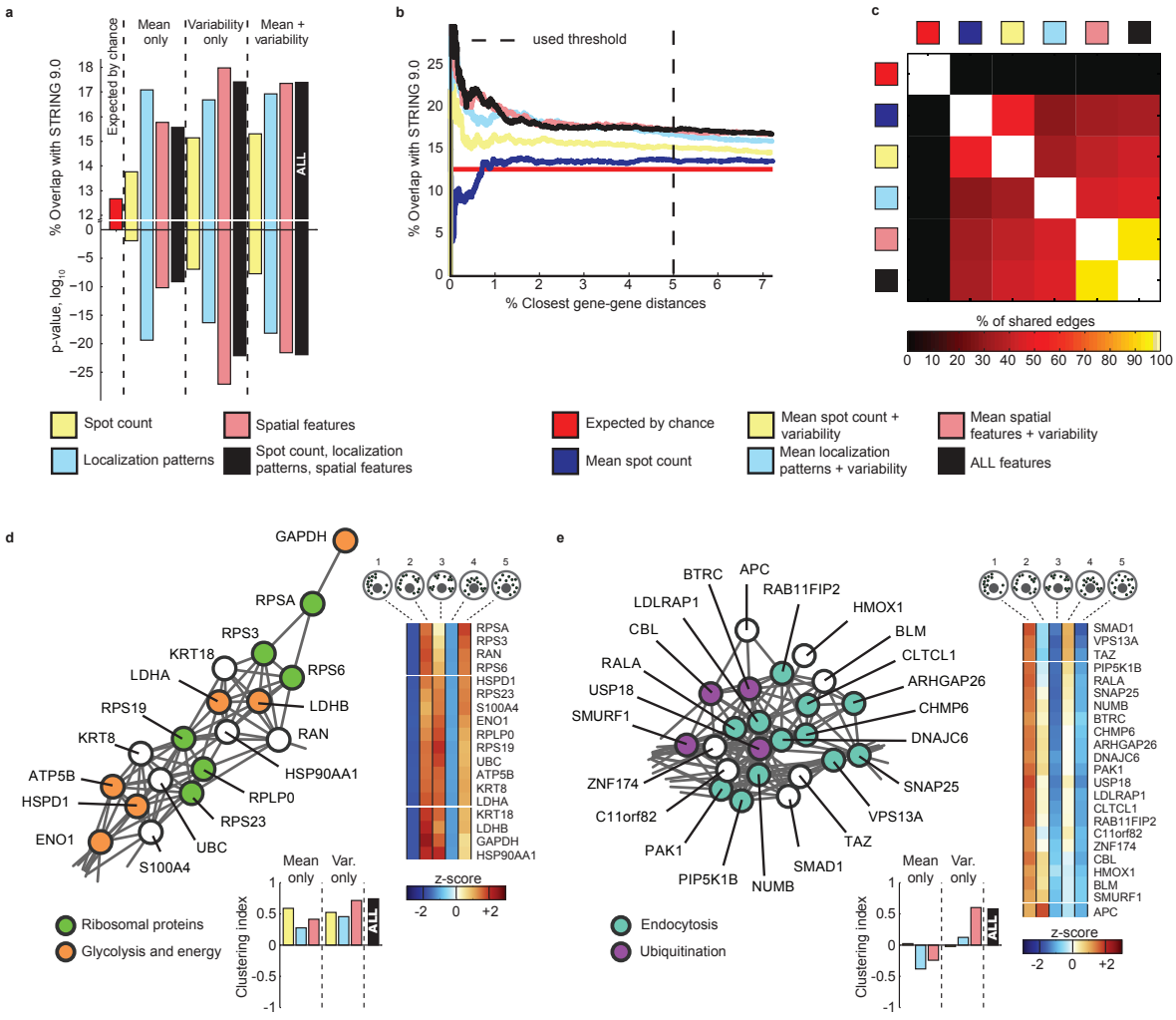


Supplementary Figure 14 | Feature selection based upon reproducibility. (**a**) Amount of features selected by individual iterations of feature selection using the feature set consisting of mean and variability features of spot count, localization patterns and spatial features (ALL features, see Fig. 5b). (**b**) Times a specific feature was selected shown for the ALL features starting set (n total = 60 selection rounds). Color boxes indicate the starting feature sets for analysis in **Fig. 5** and **Supplementary Fig. 14**: spot count (yellow), localization patterns (light blue), and spatial features (light red). Features highlighted with black dashed boxes show features describing only mean measurements. Central moments are abbreviated with Mom. (**c-d**) Fraction of genes for which the replicate pattern is within the indicated ranked distance to all genes of the replicate assay. Only genes within the testing sets of the feature selection procedure are considered (Online Methods). In **c** the reproducibility of genes using the ALL feature set before (red line) and after (black line) feature selection is shown. In **d** as in **c** but for distinct starting sets after feature selection, colors refer to **b**. Dashed lines consider only mean values of features (over both single cells and cell populations, features within black dashed boxes

Nature Methods: doi:10.1038/nmeth.2657

in **b**. Solid lines consider mean and features of the distributions of primary spot feature values (both within single cells and across single cells) (variability features). Spot count variability refers to median, variance, and 3rd to 6th central moment between cells. Localization patterns variability refers to the variance, and 3rd and 4th central moment of the classification distributions. Spatial features variability refers to the median, standard deviation, variance, and 3rd to 6th central moment between spots and to the median, variance and 3rd central moment between cells.

Nature Methods: doi:10.1038/nmeth.2657



Supplementary Figure 15 | Unbiased multivariate analysis of quantitative signatures of the *in situ* transcriptome in human tissue culture cells. (**a**) The overlap of the 5% smallest pairwise gene-gene distances [Euclidian distance between two genes based on their pairwise similarities (normalized Euclidian distance in feature space) with all other genes] with known gene interactions in STRING 9.0 and their respective p-values, calculated from various different sets of extracted features and 60 rounds of feature elimination to maximize reproducibility (Online Methods). Colors represent the type of feature sets used: 1) features of the spot count per cell (yellow); 2) features describing the types of mRNA localization patterns (light blue); 3) spatial features of spots (light red) and, 4) every feature type (black). The expected overlap by chance is also shown (red). In the first section of the bar graph only the mean values of features (over both single cells and cell populations) are considered. In the second section only features of the distributions of primary spot feature values (both within single cells and across single cells) (variability features) are considered. Spot count variability refers to median, variance, and 3rd to 6th central moment between cells. Localization patterns variability refers to the variance, and 3rd and 4th central moment of the classification distributions (Online Methods). Spatial features variability refers to the median, standard deviation, variance, and 3rd to 6th central moment between spots and to the median, variance and 3rd central moment between cells. In the third section, mean and variability features are

Nature Methods: doi:10.1038/nmeth.2657

considered. ALL (the most right black bar) indicates the set used to derive the network in **Fig. 5b**. (**b**) The overlap of pairwise gene-gene distances with known gene interactions in STRING 9.0 is shown for the 0-7% top ranking smallest distances. Colors represent the used feature sets (see legend). Dashed line indicates 5%, the threshold used for connecting two genes and further analysis in a, **c-e**, and **Fig. 5**. (**c**) Percentage of shared gene-gene distances below the 5% threshold (network edges) between various feature sets (see legend). (**d,e**) Sub-region 3 and 4 of the network in **Fig. 5b**, showing a tight cluster of genes encoding for ribosomal proteins (green) and proteins involved in glycolysis/energy metabolism (orange), and a tight cluster of genes encoding for proteins involved in endocytosis (turquoise) or ubiquitination (purple). Also depicted are z-scored mean classification distributions of cells (as clustered heatmaps) for all five main types of single-cell spot localization patterns (Online Methods) for the genes in each sub-region. Type 1 corresponds to a polarized, type 2 to a distal, type 3 to a distal and aggregated, type 4 to a perinuclear, and type 5 to a spread-out spot distribution. Bar graphs indicate the clustering index for each feature type in each sub-region (color coding as in **b** and **c**).

Nature Methods: doi:10.1038/nmeth.2657

6. Computer vision for image-based transcriptomics.

By

Thomas Stoeger*, **Nico Battich***, Markus D. Herrmann, Yauhen Yakimovich & Lucas Pelkmans.

Published in *Methods,* 01 September 2015.

doi:10.1016/j.ymeth.2015.05.016

*Contributed equally.

All algorithms described in this chapter were designed and written in equal contribution by Thomas Stoeger and Nico Battich. Markus Herrmann contributed one specific implementation of the *IdentifyPrimaryIterative* algorithm. The text of this chapter was written mainly by Thomas Stoeger, and the figures were created mainly by Nico Battich. Pseudocode was written by Yauhen Yakimovich.

Methods 85 (2015) 44–53

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Computer vision for image-based transcriptomics

Thomas Stoeger [a,b,1], Nico Battich [a,b,1], Markus D. Herrmann [a,b], Yauhen Yakimovich [a], Lucas Pelkmans [a,*]

[a] Faculty of Sciences, Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
[b] Life Science Zurich Graduate School, Ph.D. program in Systems Biology, Switzerland

ARTICLE INFO

Article history:
Received 19 February 2015
Received in revised form 13 April 2015
Accepted 17 May 2015
Available online 23 May 2015

Keywords:
Image-based transcriptomics
Single-molecule
Single-cell
Segmentation
Localization
Subcellular
High-throughput
FISH
In situ hybridization

ABSTRACT

Single-cell transcriptomics has recently emerged as one of the most promising tools for understanding the diversity of the transcriptome among single cells. Image-based transcriptomics is unique compared to other methods as it does not require conversion of RNA to cDNA prior to signal amplification and transcript quantification. Thus, its efficiency in transcript detection is unmatched by other methods. In addition, image-based transcriptomics allows the study of the spatial organization of the transcriptome in single cells at single-molecule, and, when combined with superresolution microscopy, nanometer resolution. However, in order to unlock the full power of image-based transcriptomics, robust computer vision of single molecules and cells is required. Here, we shortly discuss the setup of the experimental pipeline for image-based transcriptomics, and then describe in detail the algorithms that we developed to extract, at high-throughput, robust multivariate feature sets of transcript molecule abundance, localization and patterning in tens of thousands of single cells across the transcriptome. These computer vision algorithms and pipelines can be downloaded from: https://github.com/pelkmanslab/ImageBasedTranscriptomics.

1. Image-based transcriptomics is unique in several ways

In the past few years a wealth of techniques have been developed to study genome-wide transcriptional output at the single-cell level [1–7]. In contrast to methods relying on sequencing or PCR, image-based transcriptomics visualizes single transcripts in a population of single cells *in situ*. This allows not only the absolute quantification of transcript copy numbers, but also the spatial mapping of transcript molecules to the sub-cellular microenvironment [4]. Being an *in situ* technology, it does not require homogenization of cells and therefore minimizes the loss of material, thus achieving very high detection efficiency [4]. Another advantage of image-based transcriptomics is that it can be combined with the phenotypic characterisation of each single cell and its context within a population of cells or tissue, by microscopic assays and stainings commonly used in cell and developmental biology. This makes image-based transcriptomics of particular interest when studying the localization dynamics of the transcriptome in response to stimuli or perturbations and to identify sources of cell-to-cell variability in these processes [8,9]. While establishing image-based transcriptomics, we soon realized that a robust computer vision pipeline was as important as the experimental platform for accurately identifying and characterizing each single transcript molecule within a cell. Therefore, we here describe in detail our recent computer vision algorithms that result in accurate detection of objects in spinning disk confocal microscopy images. Besides providing a robust guide for identifying billions of individual transcript molecules with little hands-on user time, we describe how to unlock functionally important parameters of gene expression, which are impossible to grasp without the power of computer vision. For instance, multivariate descriptors of the position of each single transcript molecule enable an unsupervised characterization of the localization of transcripts of every cell.

1.1. General outline

Image-based transcriptomics employs multi-well plates to stain cells in parallel with specific probes against a transcript of interest (Fig. 1). Within single wells of a multi-well plate, the transcripts of different genes are stained by an automated experimental procedure. Each single transcript molecule is detected by high-throughput microscopy and computer vision. Experimental and computational steps can be performed with equipment that is commonly used for image-based high-throughput assays.

* Corresponding author.
E-mail address: lucas.pelkmans@imls.uzh.ch (L. Pelkmans).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ymeth.2015.05.016
1046-2023/

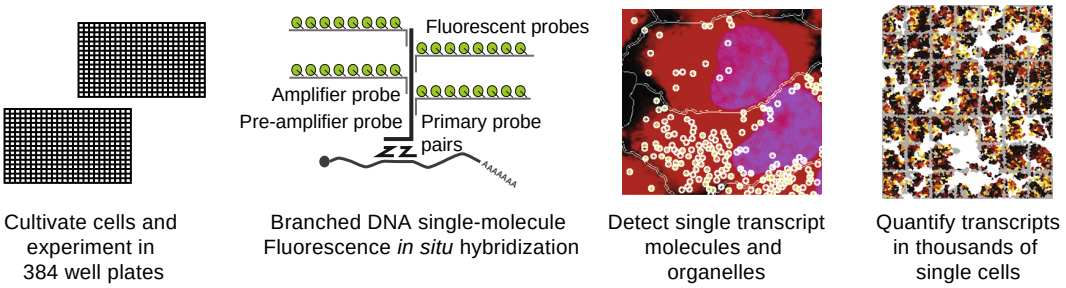


Fig. 1. Outline of image-based transcriptomics using bDNA sm-FISH.

Each single transcript molecule is stained by branched DNA single-molecule *in situ* hybridization (bDNA sm-FISH). This technology, which is commercially available from Affymetrix and Advanced Cell Diagnostics, applies a series of consecutive *in situ* hybridizations, which visualize each single transcript molecule as a bright fluorescent spot. In a first round of *in situ* hybridization, two epitope-specific primary probes bind next to each other on the same transcript molecule. While it is technically possible to implement bDNA FISH with only one epitope-specific probe [10], requiring the simultaneous binding of two probes in direct spatial adjacency should reduce unspecific signal [11]. Targeting 15 different epitopes of each transcript in a single hybridization reaction ensures that at least one epitope is accessible to the detection reagents without the need to denature the specimen. The subsequent rounds of *in situ* hybridization create a docking platform for ~500 fluorescently labelled probes per single epitope. This level of fluorescence is sufficiently high to enable the specific, rapid and robust detection of single transcript molecules by high-throughput imaging.

1.2. Alternative methods for RNA detection in imaging

Another method for directly visualizing single transcript molecules *in situ* is oligonucleotide-based single molecule FISH (o-nuc sm-FISH). This approach targets individual transcripts by up to 40 different oligonucleotides, which are directly conjugated to fluorophores. While a recent study achieved to monitor 61 different ncRNAs, it had to restrict itself to "a few dozen cells ... due to limited imaging throughput" [12]. Possibly, this reflects the lower signal-to-noise ratio of single fluorescent spots of o-nuc sm-FISH and their need for a 600 times longer illumination time [4].

Alternatively, transcripts can be visualized indirectly via reverse transcription to cDNA that can be sequenced *in situ* by padlock probes [13] or oligonucleotide ligation and detection [14,15]. While the former sequencing approach can presently detect 31 different genes simultaneously in thousands of single cells within a tissue slide [13], the latter approach can currently read around 200 mRNAs simultaneously for 40 different cells [15]. The efficiency for detecting single transcript molecules has been estimated to be 30% [13,16] and 3% [15] respectively, which is much lower than the 85% of hybridization efficiency in sm-FISH [4,17]. Such low efficiencies currently prevent these alternative methods from surveying the transcriptome with single-molecule sensitivity and resolution *in situ* [18,19].

2. Establishing image-based transcriptomics with single molecule resolution

The detailed experimental protocol for high-throughput bDNA sm-FISH has been published previously [4] and therefore, we here mainly provide additional assistance for setting up a robust automated experimental platform. As a general introduction to high-throughput image-based assays and the infrastructure and software supporting such experiments we highly recommend the excellent essay by Buchser and colleagues [20].

Table 1 contains an overview of potential problems occurring during the detection of single transcripts. The most critical factor in getting reliable results is to use an automated incubator that contains rotating towers for the individual storing of multi-well plates during hybridization reactions. This prevents the occurrence of different hybridization efficiencies in different wells of a multi-well plate (data not shown). Table 2 highlights potential pitfalls, which could affect the biological interpretation of accurate single-molecule measurements. We recommend repeating the control experiments suggested in Table 1 and Table 2 in different weeks to ensure that your setup of image-based transcriptomics functions robustly.

3. Establishing the image analysis pipeline

A robust image analysis pipeline is required for accurate measurements of absolute transcript levels as well as measurements of transcript localization in the cytoplasm of single cells, and extraction of features that describe the cellular phenotype. First, homogeneous intensity values throughout the images in all channels must be ensured, and then object segmentation must be performed minimizing errors. To ensure this, we developed four algorithms to perform high-throughput illumination correction of raw images, robust nuclei and cell segmentation, and robust spot detection. They can be downloaded from https://github.com/pelkmanslab/ImageBasedTranscriptomics and applied on an example dataset available on https://image-based-transcriptomics.org. The algorithms presented in this manuscript do not intend to replace single-cell quality control. For the latter we recommend interactive user-guided supervised machine learning, which has been implemented before by our group [23] and others [24]. Supervised machine learning not only readily identifies rare cells that have not been correctly segmented, but also allows the selection of a group of cellular objects that is relevant for a specific biological question (e.g., interphase cells).

The algorithms presented in this manuscript intend to reduce human hands on time and increase the amount of high-quality primary data after computational image-analysis (Table 3). Computational running time has not emerged as a practical issue for image-based transcriptomics. The algorithms are robust in the sense that their input parameters rarely have to be adjusted for individual experimental plates.

While the principles of the algorithms presented in this manuscript have been sketched in one of our earlier publications, the description beneath provide a detailed guide for using those algorithms. Moreover we here include implementations of these

Table 1
Suggested controls for the detection of transcript molecules.

Possible artefact	Experiment	Hints
Inability to detect single molecules	Assay with probes against a single epitope of HPRT1 [4]	Exposure time during imaging; protease concentration
False positive detection	Probe against bacterial gene dapB. Less than 1 spot per cell should be detected	Protease concentration; cells without cytoplasmic DNA
Spill-overs	Stain adjacent wells for the negative control (bacterial dapB) and the highly abundant ACTB transcripts. Test full plates	Liquid handling
Efficiency of single molecule detection	Stain same transcript on two different sets of epitopes by two different sets of amplification reagents, which can be visualized by two different fluorophores. Efficiency of detection should be ~85% [4]	Protease concentration; amount of targeted epitopes per transcript; computational spot detection
Positional bias between wells (staining reaction)	Stain all wells of a multi-well plate with probes against the non-abundant housekeeping gene HPRT1	Always use incubator with rotating towers for hybridization; never skip protocolled in-solution mixing
Low reproducibility	Multiple independent assays across different weeks	Aberration of liquid handling < 1%; cell seeding
Tearing of signal of single molecules	Compare signal obtained by multiple units and types of objectives	Choose best objective and remove remaining effect computationally (see below)

algorithms for MATLAB and implementations as modules for CellProfiler to segment single nuclei and cells.

3.1. Illumination correction

Illumination correction of raw images is essential for subsequent steps in the image analysis pipeline. It ensures correct object detection and accurate measurements of intensity features, reducing biases due to uneven illumination of the sample as well as positional differences in the signal gain resulting from the detection system. During image-based transcriptomics, we exploit the statistical power of the large number of images acquired per channel to learn pixel-wise illumination and signal gain biases (Battich et al. [4], Fig 2). Briefly, we calculate the standard deviation and mean intensity values per pixel for every pixel position of a given channel. To correct the illumination bias, per-pixel *z*-scoring is performed as shown in Fig. 2(Eq. (1)). The *z*-scored values are then reversed to intensity values as shown in Fig. 2(Eq. (2)).

3.2. Nucleus segmentation

Pixels belonging to nuclei objects can be easily distinguished from background pixels by thresholding an image of a nuclei-specific stain such as DAPI. However, this often results in clumps of several nuclei because a single, image-wide threshold value is generally not sufficient to separate nuclei that lie very close to each other. Such clumped objects are relatively large and display multiple concave regions. Generally, at the intersection of individual objects, a line of low intensity pixels connects two concave regions, which can be found by the watershed algorithm [25]. Thus, we identify single nuclei with an algorithm consisting of two parts: first, intensity thresholding by the Otsu method [26] identifies primary objects; and, secondly, objects consisting of multiple nuclei are separated along the best identified watershed line (Fig. 3).

The algorithm (algorithm 1) uses illumination-corrected images and processes them as follows:

1) Initial objects are identified by simple thresholding.
2) Clumped objects are selected on the basis of size and shape features: area, solidity, and form factor.
3) The perimeter of selected objects is analysed and concave regions along the boundary of objects are identified.
4) Putative watershed lines connecting two concave regions are determined using the Dijkstra shortest-path algorithm [27].

Table 2
Suggested controls for the proper interpretation of single molecule measurements.

Possible artefact	Experiment	Hints
Positional bias between wells (biological)	Compare the number of cells and local cell density [21] across individual wells	Avoid "edge-effects" by following the cell seeding protocol of Lundholt et al. [22]
Variable number of cells per seed	Monitor and potentially adjust cell dissociation protocol such that, on average, a cell aggregation score of less than 1.2 is achieved for each seed	Trypsinization time; repeatedly shear cells through pipette pressed towards plastic dish
Loss of cells during assay	Perform live-imaging of cells with Hoechst dye prior to the assay and compare with presence of cells after image-based transcriptomics	Slowly pipet to side of well (most steps) or center of well (only for in-solution mixing)
Staining of cell-outline varies between experiments	Repeat and time succinimidyl ester staining with multiple freshly prepared staining solutions	Time-dependent decay of carboxylic acid, succinimidyl ester, in aqueous solutions

Table 3
Time for image-based transcriptomics on ten 384-well plates to obtain results, whose quality appeared acceptable to us. Time estimates are based on our experience and depend upon the specific computational infrastructure.

	Hands-on time	Computational time	Computers used by us
Illumination correction (this manuscript)	5 min	2–5 h	4
Nucleus segmentation (CellProfiler)	30 min	<<1 h	1500
Nucleus segmentation (this manuscript)	30 min	<<1 h	1500
Cell segmentation (CellProfiler)	5–10 h	<<1 h	1500
Cell segmentation (this manuscript)	5 min	2–10 h	1500
Cell segmentation (manual segmentation)	>> 1 month (expected)		
Inference of spot detection parameters (this manuscript)	1–2 h	2–8 h	10
Optional lens aberration correction	1 h	1–2 h	1500
Spot detection (this manuscript)	30 min	<<1 h	1500
Measuring localization of transcripts	5 min	1–2 h	1500

5) All possible cuts along the selected watershed lines are considered and features of each potential cutting line (intensity along the line, angle between concave regions) as well as features of the resulting objects (area/shape) are measured.
6) An "optimal" cut line is finally chosen by minimizing a cost function that evaluates the measured features. The resulting objects have a minimal size and are as round as possible, while the separating line is as straight and as short as possible and the intensity along the line as low as possible.

Algorithm 1 IdentifyPrimaryIterative

```
1.  Initialize() // initialize objects by thresholding the input intensity image;
2.  InitialSegmentation() // recognize objects as segmented objects without cutting them first;
3.  Repeat
4.    For each object in segmented objects
5.      If lower size threshold < size of object < upper size threshold
6.          and solidity of object < solidity threshold
7.          and transformed form factor of object > form factor threshold
8.      then
9           Add object to the collection of clumped objects to be cut;
10.     end
11. End
12. PerimeterAnalysis() // analyze perimeter of selected clumped objects and calculate the curvature along their boundary [see PerimeterAnalysis.m];
13. PerimeterWatershedSegmentation() // cut selected clumped objects along watershed lines between concave regions [see PerimeterWatershedSegmentation.m];
14. For each object in clumped objects
15.     IdentifyConcaveRegions() of the object, where region is concave
16.     If angle of circle segment of region > equivalent segment threshold
17.     and radius of region < equivalent radius threshold
18.     IdentifyLinesAndNodes() of the object // find all watershed lines and nodes, where each node is a single pixel on the line that overlaps with the object boundary;
19.     Select watershed nodes If node lies within concave regions;
20.     Select all watershed lines If line connects two watershed nodes from different concave regions;
21.     For each line in watershed lines
22.       Measure line length and straightness and the intensity profile along the line;
23.       Measure the angle between normal vectors at watershed nodes;
24.       Measure area, solidity and form factor of the cut object, i.e. the smaller of the two objects that would result from a cut along the line.
25.       If size of the object < threshold of object being too small
26.       then
            discard such cutting line from selected watershed lines;
27.       end
28.       Select "optimal" watershed line by minimizing the cost function
```

Algorithm (*continued*)

Algorithm 1 IdentifyPrimaryIterative

```
29.       cost (a, b, c, d, e, f, g, h) ← a−2 ∗ b − c − d − e + 2 ∗ f − g − 2 ∗ h,
30.       where
            a is a solidity of cut object,
            b is a form factor of cut object,
            c is a mean intensity along the line,
            d is a max intensity along the line,
            e is a 0.75 quantile intensity along the line,
            f is an angle between two watershed nodes,
            g is a line straightness,
            h is a line length.
31.     End // of for-each-loop at line 21
32.   End // of for-each-loop at line 14
33. Until no more clumped objects can be found.
```

Whenever attempting to identify individual nuclei of a novel cell line or whenever changing imaging conditions, we recommend to empirically test different schemes and parameters for segmentation of nuclei. Good settings can usually be found empirically by using the inbuilt testing mode of IdentifyPrimaryIterative. Contrasting CellProfiler's default options for separating objects, which are part of the IdentifyPrimaryAutomatic module, IdentifyPrimaryIterative can simultaneously consider the local intensity of the DAPI stain and the geometry of identified objects to separate them. In practice we never had to adjust the threshold value suggested by the Otsu method [26]. Depending on the biological question of interest, one might choose settings for the separation of objects, which favour over- or under-segmentation. For instance over-segmentation increases the fraction of emerging cells during anaphase cells that are already considered as individual objects. Under-segmentation on the other hand facilitates the correct segmentation and thus quantification of multinucleate cells.

Frequently not every object, which can be identified in image-based assays, should be considered in subsequent analysis. For instance we preclude the analysis of DAPI positive cellular debris, apoptotic bodies and mitotic cells by a dual strategy, which is independent of IdentifyPrimaryIterative. First the DiscardObjectsBySize.m module removes small objects within CellProfiler. Second, supervised machine learning identifies debris, and apoptotic and mitotic cells [23].

3.3. Cell segmentation

The segmentation of cells uses the segmentation of nuclei as seeds [28]. It is imperative to ensure correct segmentation of the cellular cytoplasm as this will not only have a major impact in the number of spots (or transcript molecules) allocated to each cell, but will also drastically affect measurements of transcript localization. To achieve the high accuracy in cell segmentation required for image-based transcriptomics, we developed an algorithm that performs sequential rounds of watershedding, rather than the one round of watershedding typically applied [28]. This iterative algorithm allows accurate identification of the boundary between cells with relatively minimal user input.

In the algorithm, an arbitrary amount of different segmentations are combined in such a way that the allocation of single pixels to their correct seeds (nuclei) never becomes worse and thus becomes optimal by iteratively performing many different segmentations (Fig. 4). Besides largely eliminating human hands-on time, this strategy generally yields superior results compared to a single segmentation: different parts of a single cell can be segmented by opposing segmentation settings, which only yield

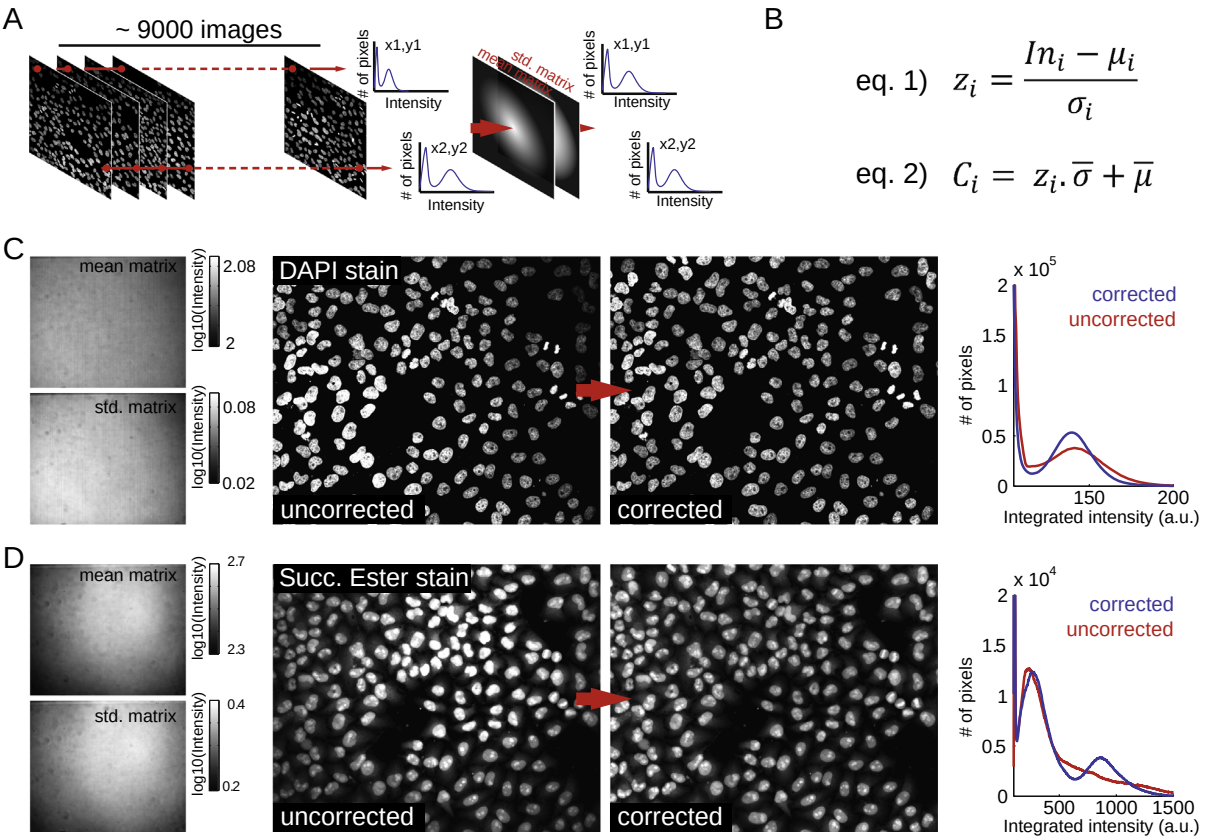


Fig. 2. (A and B) Method for illumination correction of images. For each channel the mean intensity μ_i the standard deviation σ_i are calculated for each pixel p_i in the field of view. Then an overall mean intensity $\bar{\mu}$ as well as the mean standard deviation $\bar{\sigma}$ of all pixels is derived from the "mean" and "std." matrices. Illumination correction is performed by per-pixel z-scoring (Eq. (1)), where z_i the z-scored value for pixel p_i and In_i is the original intensity value for pixel p_i in a given image. The corrected intensity value C_i for pixel p_i in an image was then calculated as in Eq. (2). C) Illumination correction examples for the DAPI channel. D) As in C but for Alexa Fluor succinimidyl ester (a general protein stain).

optimal segmentation accuracy in a subpart of the cell, but perform sub-optimally in other subparts.

Briefly, the algorithm (Algorithm 2) treats the input images as follows:

1) Calculate the watershed cell segmentation at different thresholds.
2) One label image is constructed. If a pixel is part of different objects at a given threshold (which is likely in cell-rich regions), it will be allocated to the object of the higher threshold (e.g. if threshold specifications were 1 and 0.5, it would be attributed to the object identified with a threshold of 1).
3) Define background pixels by a single user-provided microscope-specific threshold, which can be determined manually once.
4) Re-label pixels of prospective objects (cells), which are not connected to their original seed (nucleus), as pixels belonging to the background.

Algorithm 2 IdentifySecondaryIterative

1. 1. Initialize empty *FinalSegmentation* matrix;
2. Load *OrigInputImage* matrix;*DefineThresholds*(*OrigInputImage*) // defines a sorted list of *thresholds* $\{T_i\}$, where T_{min} is a minimalthreshold [e.g. chosen by CPthreshold.m] and $T_i < T_{i+1}$.
3. *SeedMarkersImage* ← *DefineSeedObjects*(*OrigInputImage*) // labels each pixel with grayscale values,uniquely enumerating each *seed object*(e.g. nuclei) by its ID (usually, a foreground mask from previous segmentation);
4. **For each** *threshold* T_i **in** *thresholds* $\{T_i\}$
5. Obtain binary *ThresholdedImage* of pixels, **where** each *pixel* = 1
   If pixel intensity > T_i **and** *pixel* **not in** *foreground*
   else *pixel* = 0;
6. Find labeled segmentation S_i ← *WatershedMethod*(*ThresholdedImage*, *SobelGradient*(*OrigInputImage*), SeedMarkersImage); // [see *IdentifySecondary.m for* '*Watershed*' choice of method];
7. **Select** *indexes* of all non-background pixels **in** segmentation S_i;
8. *FinalSegmentation*(*indexes*) ← S_i (*indexes*); // overwrite selected pixels within FinalSegmentation with the label values in S_i;

Algorithm (*continued*)

Algorithm 2 IdentifySecondaryIterative

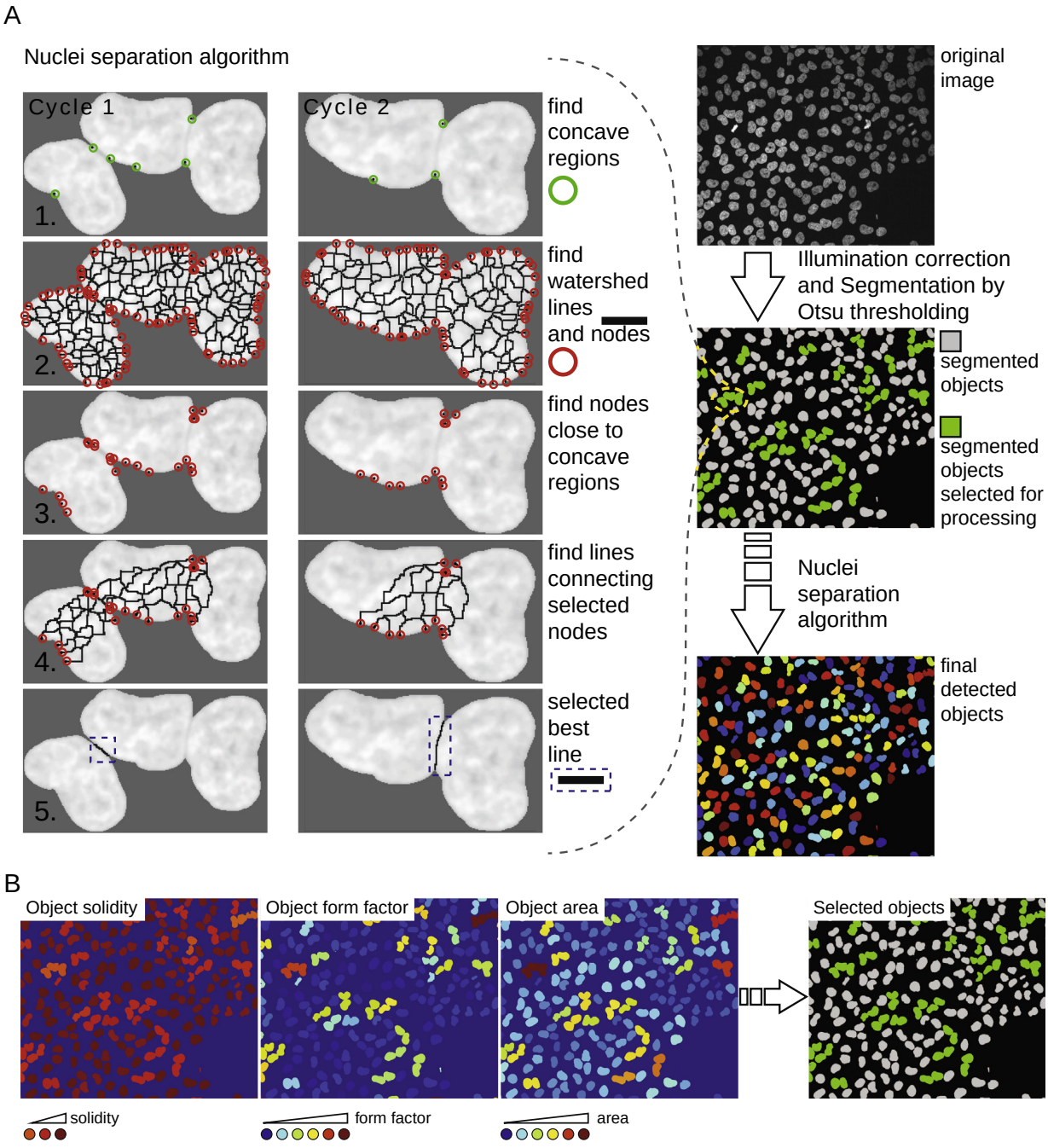


Fig. 3. (A) Scheme for nuclei segmentation and iterative correction of primary segmented objects. (B) Strategy for selection of objects to be separated by combining the object solidity, form factor and area. All features measured as in the CellProfiler module "MeasureObjectAreaShape.m".

Algorithm (*continued*)

Algorithm 2 IdentifySecondaryIterative

9. **End**
10. *CleanSegmentation*(*FinalSegmentation*) // Make sure *FinalSegmentation* complies to CellProfiler expectations.

Function *CleanSegmentation*(*FinalSegmentation*):

1. Identify borders between different object labels as the non-zero values upon Sobel-filtering, i.e. *SobelGradient*(*FinalSegmentation*);
2. Set the identified borders between different label values of objects to background;
3. Set pixels with the object labels to background value, if other pixels with the same object labels do not connect them to the seed with the same label values of objects.

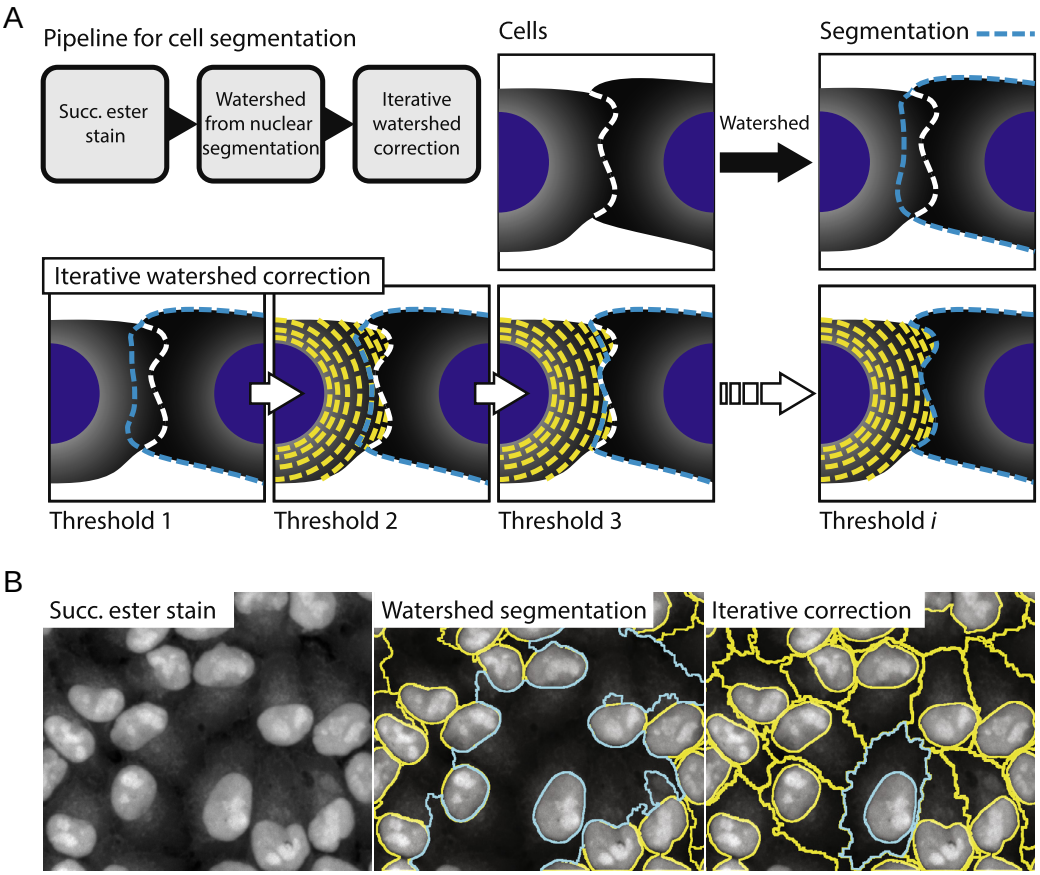


Fig. 4. Improvement of segmentation of cells by iterative correction. Several different and partially overlapping segmentations are combined to a single optimal segmentation (Panel A). Detection of single cells stained by carboxylic acid, succinimidyl ester (Panel B).

In our experience IdentifySecondaryIterative has never been performing worse than segmentation by a single round of wathershedding. The few remaining miss-segmented cells can be identified by supervised machine learning. As with any image-based assay the ability to resolve fine structures of the cellular periphery depends upon their size and the resolution of the microscopic images. Like other algorithms that segment 2D images to segment cells, IdentifySecondaryIterative works best on cells that do not grow on top of each other, such as HeLa Kyoto cells, RPEI cells and primary keratinocytes. If cells can grow on top of each other, it is not always possible to allocate a single pixel to a single cell (e.g.: A431, NIH 3T3, HEK293), though supervised machine learning could be used to discard those cells, which grow on top of each other.

3.4. Spot detection and correction of lens aberrations

The basic strategy for detecting single transcripts as spots has been developed by Jiri Matas [32] and Arjun Raj and their colleagues [17]. After emphasizing spot-like signal by a Laplacian of Gaussian filter (Fig. 5A), a threshold for the detection of objects is chosen such that, on each single image, the specific value of the threshold only mildly affects the number of detected transcripts (Fig. 5B and C). As the numerical value of the threshold will partially depend upon the absolute intensity of the acquired images, we rescale the intensities of individual images such that they are comparable between different images and a single numerical value for the threshold can be chosen (Fig. 5C). This seemingly minor, but essential, detail of our image-analysis pipeline contrasts the most common high-throughput implementation of spot detection algorithms, which rescales the intensities of any image according to the intensities of its dimmest and brightest pixel [17,28,33]. While the accompanying code supports additional refinement of the spot detection, these additional parameters (2D/3D, minimal intensity of pixels, size of spots) have a negligible effect on the detection of transcripts once robust imaging conditions have been established experimentally.

For identifying the settings for detecting spots, we include on each experimental plate 4 wells in which bacterial dapB transcripts are probed (a negative control for mammalian cells), and 4 wells for probing transcripts of the housekeeping gene HPRT1 (Hypoxanthine Phosphoribosyltransferase 1, which plays a central role in purine nucleotide synthesis). In a first computational step, we find the upper- and lower-image intensity boundaries for those two reference transcripts (see Exp_getIntensitiesOfReferences.m). The next step detects spots at varying threshold values while rescaling the intensities of single images according to the previously identified bounds (see Exp_getSpotCountsOfReferences.m). Upon completion of the computation, a threshold is chosen manually such that its specific numerical value only mildly affects the number of detected transcripts (see Exp_selectDetectionThreshold.m). In practice, a fast manual choice and optimization of settings is as good as a fully computational procedure, but offers the advantage of being a first quality control of the data. The number of spots in the dapB negative control should be much lower and

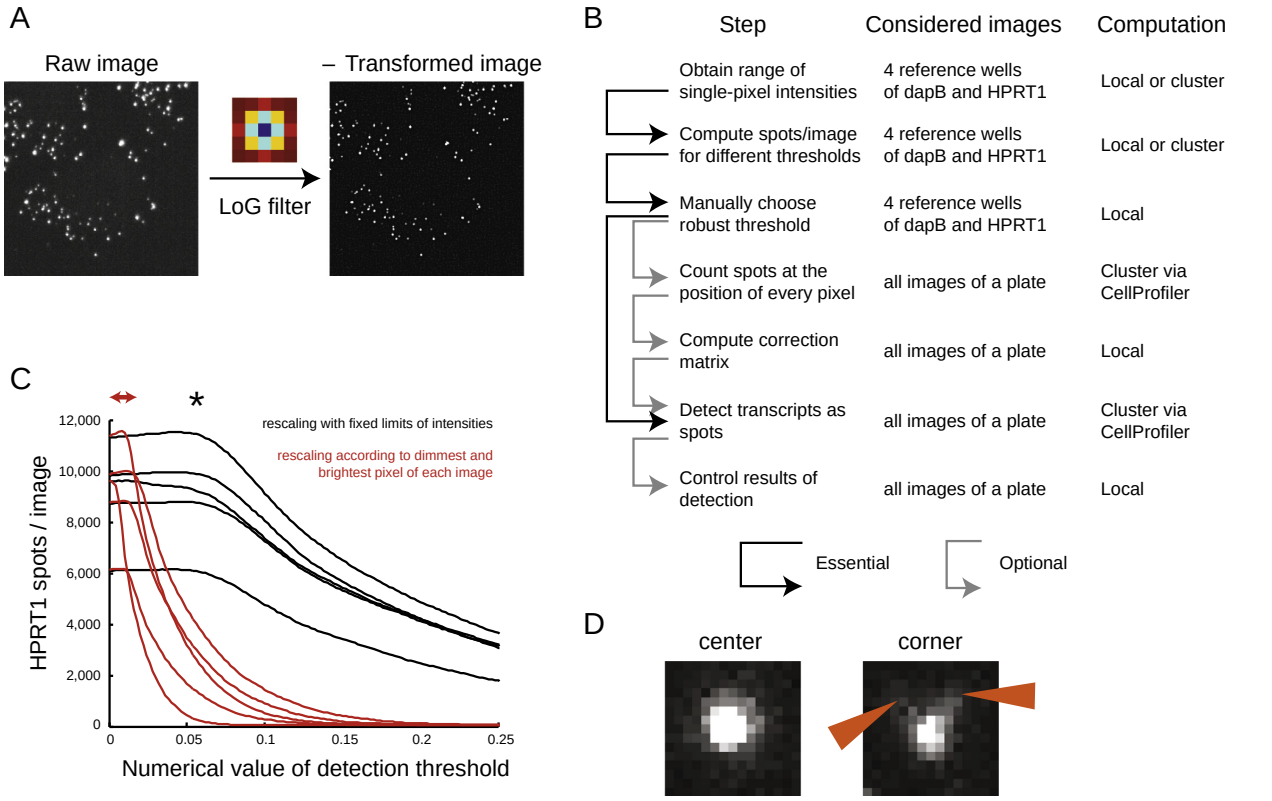


Fig. 5. Detection of single transcripts as spots. Application of a Laplacian of Gaussian (LoG) filter emphasizes spot-like signals (Panel A). Workflow for detecting transcripts as spots (Panel B). The specific numerical value for the detection threshold only mildly affects the number of spots once the intensities of individual images are rescaled similarly. Lines represent five different, randomly chosen images; arrows and asterisk indicate suggested thresholds (Panel C). The signal of individual transcripts is slightly teared in the corner of an image (Panel D).

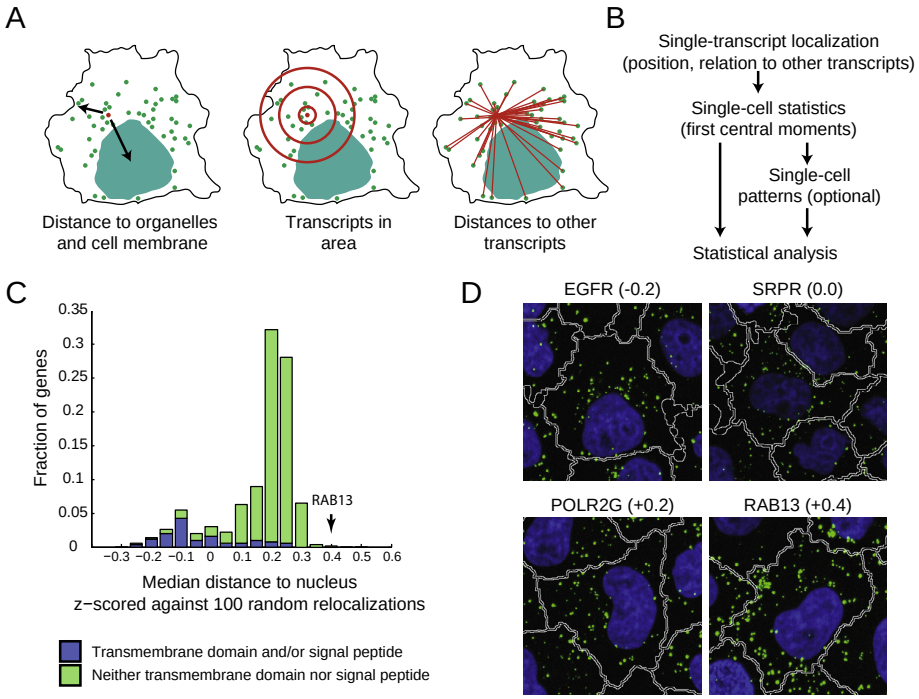


Fig. 6. Readouts of single-transcript localization (Panel A). Pipeline of converting single-transcript readouts to single-cell readouts (Panel B). Inspecting expected behavior of basic measurements of the localization of transcripts. The distance of transcripts to the nucleus is shorter for transcripts translated at the ER (Panel C). Median distance of all transcripts is normalized by *z*-scoring against 100 relocalizations of the transcripts to random pixels of the cytoplasm. Median of all cells over all single-cell medians is shown (Panel D). Differing distances to the nucleus become apparent to humans in large cells upon visualizing transcripts (green), the nucleus (blue) and the cell outline (white lines) (Panel D, numbers as in Panel C).

more sensitive towards changes in the numerical value of the threshold [4].

Optical aberrations, which tear the signal of individual transcript molecules in the corners of an image, make the signal less spot-shaped. This creates a spatial bias in the detection of transcript molecules of approximately ± 3% at different positions of an image [4]. While it is best to reduce this effect experimentally (see above), it can be optionally attenuated further by computationally modifying the threshold for the spot detection at different positions of an image. Use the ScanSpotThresholds.m CellProfiler module to test multiple different thresholds surrounding the previously identified reference threshold. Inclusion of all images of a plate (recommended: approximately 10,000 images), allows computing the spatial bias of the detection of spots, which can be used to construct a correction matrix that will modify the spot detection threshold for each pixel (see Exp_computeCorrectionMatrix.m).

You can now identify spots with a CellProfiler pipeline containing the IdentifySpots2D.m module; optionally apply a correction matrix against the spatial bias, which can be loaded by the LoadSingleMatrix.m module; and, insert the parameters for the spot detection determined above. Additionally, we recommend enabling the deblending option, an algorithm from astrophysics [34], which can spatially resolve individual transcript molecules below the optical diffraction limit. If a correction matrix for the spatial bias has been applied, monitor its impact on the spatial bias of the spot detection (see Exp_checkBiasCorrection.m) and potentially restrict or expand the range of thresholds that have been considered for the construction of the correction matrix.

In addition to the algorithm outlined above, which provides highly reproducible and specific measurements of the number of transcripts in a high-throughput experimental setup with bDNA sm-FISH [4], we would like to note several excellent algorithms, that have been used with o-nuc sm-FISH to identify those fluorescent spots that likely indicate single transcripts [29–31].

3.5. Quantification of spot localization

Being an *in situ* technology, image-based transcriptomics can quantify the localization of each single transcript molecule. Although the subcellular localization of transcripts and it variability across single cells can hold more biological information than single-cell transcript abundance [4], it is not yet used routinely in functional genomics studies due to technical limitations. This section describes how this powerful source of information can be unlocked from image-based transcriptomics data.

Each single transcript molecule can be characterized by a set of measurements (Algorithm 3), which describe its distance to the centroid or edge of an organelle or the cell [4]. In addition, the position of each transcript molecule can be placed in relation to other molecules, for instance by measuring the variance of its pairwise distances to all other molecules, or by counting the number of transcript molecules within a certain area. Such readouts of single molecules are created by the MeasureLocalizationOfSpots.m CellProfiler module [4]. By choosing an arbitrary amount of differently sized areas, different scales of subcellular crowding can be compared.

Algorithm 3 CPgetSpotLocalizations(LookupImage, VectorWithDistancesForFractions, VectorWithDistanceContainingFractionOfSpots)

Initialize *Results* // a key-value array, containing all measurements;
Define *SpotDistances* as all Euclidean distances between spot pairs (cartesian product);
// Determine fractions of spots within given distance and distances for given fractions of spots;
For each *spot* **in** *spots*
For each *DistanceOfFraction* **in** *VectorWithDistancesForFractions*
Results[*FractionOfSpotsAtDistance*] ← *normalize* over
    Select all *spots* within given *DistanceOfFraction* **exclude** the *spot* itself;
    End
    For each *DistanceContainingFractionOfSpots* **in** *VectorWithDistanceContainingFractionOfSpots*
Results[*DistanceContainingFractionOfSpots*] ← **Select** min(
all *SpotDistances* for a given *spot* **within** *DistanceContainingFractionOfSpots* **exclude** the *spot* itself);
    End
End
Results[*MeanDistance*] ← *mean*(columns of *SpotDistances*);
Results[*VarianceDistance*] ← *variance*(columns of *SpotDistances*);
Results[*StandardDeviationDistance*] ← *sqrt*(*ResultsVarianceDistance*);
Results[*DistanceToCellCentroid*] ← measure distances of all *spots* to centroid of the cell;
// Treat spots at the cell membrane specially.
For each *spot* **in** *spots*
    Determine coordinate of the *closest membrane pixel*;
    Results[*ShortestDistanceToMembrane*] ← *EuclidianDistance*(*centroid of spot, closest membrane pixel*);
Results[*DistanceToNucleus*] ← *EuclidianDistance*(*centroid of spot, centroid of nucleus*);
    If *EuclidianDistance*(*centroid of spot, closest membrane pixel*) > *sqrt(2)* **then**//spot is not at the membrane;Construct a projection line connecting the *centroid of the nucleus* and *centroid of the spot*;
    Results[*DistanceAlongProjection*] ← *EuclidianDistance*(*centroid of spot, closest membrane pixel*) // membrane pixel is picked along the projection line;
    else
      Results[*DistanceAlongProjection*] ← *ResultsShortestDistanceToMembrane*]
  end
    End
Results[*MembraneBorderingCell*] ← look up pixel at position within *LookupImage* // *LookupImage* is an image indicating for each pixel, whether closest membrane is adjacent to a cell);
return *Results*.

Cellular readouts of transcript localization can be derived from the readouts of single transcript molecules. For instance, one may compute the first central moments of the distribution of every readout across all single transcript molecules within a single cell with the accompanying MeasureChildren.m CellProfiler module [4], and subsequently quantify properties of the single-cell distributions of these central moments. In practice, these information-rich multivariate readouts for each single cell, generated for thousands of cells in a single population, rarely lend themselves to ready interpretation or presentation. Therefore, we have previously developed and documented [4] an unsupervised clustering scheme that uses selected cellular statistics to identify a small number of main patterns in single cell subcellular transcript localization. This analysis has been well described by us [4] and can be computed independently of CellProfiler by our locpatterns package (https://github.com/pelkmanslab/locpatterns). Briefly, this package uses the per-cell mean and standard deviation of the single-transcript localization features to first identify a number of different patterns, by clustering random subsets of cells, such that the clusters are most reproducible. In a second step, it determines the similarity of each single cell to each of the identified patterns.

Supervised machine learning can be further applied to classify cells with a distinctive subcellular localization of transcripts [23].

One convenient way to evaluate the basic computational quantification of the localization of transcript molecules is the median distance of all transcript molecules to the nucleus. Plotting the median of this single-cell readout for multiple genes should yield a bimodal distribution (Fig. 6A): transcripts, which become translated at the endoplasmic reticulum (ER), should have a shorter distance to the nucleus compared to transcripts with a cytoplasmic translation. For instance, we noticed that transcripts of RAB13, which have previously been described to enrich in filopodia [35], tended to be furthest from the nucleus (Fig. 6B). One way of controlling finer details of the localization of transcripts is the unbiased clustering of genes by multiple readouts of the localization of transcripts. Mitochondrially-encoded transcripts should be identified as a group of colocalizing transcripts even when mitochondria are not stained [4]. Furthermore, at least in HeLa cells, one should observe a further sub-clustering of different groups of mitochondrially-encoded transcripts reflecting different positions within the mitochondria [4]. In addition, this may reveal further subclustering of transcripts translated at the ER [4], as well as transcripts translated in the cytoplasm. Such findings indicate extensive functional subcompartmentalization of the transcriptome, both on organelles and in the cytoplasm, which are properties of posttranscriptional control of gene expression that have remained hidden thus far.

4. Conclusion

Image-based transcriptomics combines precise counting of transcript molecules with a unique multivariate quantification of the subcellular position of each single transcript molecule for thousands of genes in tens of thousands of single cells. Being an image-based *in situ* technology it can be readily combined with image-based assays, which monitor additional specific biological markers of interest. To enable such lines of research, every experimental and computational step of image-based transcriptomics needs to be highly reproducible across different weeks and geared towards the quantification of single molecules. To enable image-based transcriptomics to reach its full potential, we developed computer vision algorithms that build on and improve those currently used to detect objects in confocal images. By using iterative watershedding we have improved the segmentations of nuclei and cells. In addition, we describe how to perform spot detection for transcript identification in an automated way for thousands of images. Accurate detection of nuclear outlines, cell outlines, and transcript molecules are essential for the correct quantification of a high-dimensional multivariate feature space of each transcript and to reveal bona fide novel properties of the spatial organization of the transcriptome [4]. The computer vision pipeline presented here complements our earlier work [4], and can be used independently of transcripts in other image-based approaches. It also forms a practical guide on how to extend image-based-assays to mapping small particles relative to spatial hallmarks of single cells. Indeed, the highly robust and automated protocol of the underlying computer vision pipeline has been instrumental for uncovering parameters of gene expression, which remain otherwise hidden.

Acknowledgments

We would like to acknowledge A. Schwab for help on the development of the IdentifyPrimaryIterative.m module, Q. Nguyen and S. Lai from Affymetrix for helpful comments on experimental procedures, and V. Green for useful comments on the manuscript. L.P. acknowledges financial support for this project from the Swiss National Science Foundation, the University of Zurich and the University of Zurich Research Priority Program in Systems Biology and Functional Genomics.

References

[1] F. Tang et al., Nat. Methods 6 (2009) 377–382.
[2] S. Islam et al., Genome Res. 21 (2011) 1160–1167.
[3] T. Hashimshony, F. Wagner, N. Sher, I. Yanai, Cell Rep. 2 (2012) 666–673.
[4] N. Battich, T. Stoeger, L. Pelkmans, Nat. Methods 10 (2013) 1127–1133.
[5] S. Picelli et al., Nat. Protoc. 9 (2014) 171–181.
[6] A.R. Wu et al., Nat. Methods 11 (2014) 41–46.
[7] H.C. Fan, G.K. Fu, S.P. Fodor, Science 347 (2015) 1258367.
[8] P. Liberali, B. Snijder, L. Pelkmans, Nat. Rev. Genet. 16 (2015) 18–32.
[9] N. Crosetto, M. Bienko, A. van Oudenaarden, Nat. Rev. Genet. 16 (2015) 57–66.
[10] J.R. Sinnamon, K. Czaplinski, Methods Mol. Biol. 1206 (2015) 137–148.
[11] F. Wang et al., J. Mol. Diagn. JMD 14 (2012) 22–29.
[12] M.N. Cabili, Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution, Genome Biol. (2015).
[13] R. Ke et al., Nat. Methods 10 (2013) 857–860.
[14] J.H. Lee et al., Science 343 (2014) 1360–1363.
[15] J.H. Lee et al., Nat. Protoc. 10 (2015) 442–458.
[16] C. Larsson, I. Grundberg, O. Soderberg, M. Nilsson, Nat. Methods 7 (2010) 395–397.
[17] A. Raj, P. van den Bogaard, S.A. Rifkin, A. van Oudenaarden, S. Tyagi, Nat. Methods 5 (2008) 877–879.
[18] E. Shapiro, T. Biezuner, S. Linnarsson, Nat. Rev. Genet. 14 (2013) 618–630.
[19] A.K. Shalek et al., Nature 510 (2014) 363–369.
[20] Buchser, W. et al. in Assay Guidance Manual (eds G. S. Sittampalam et al.) (2004).
[21] B. Snijder et al., Nature 461 (2009) 520–523.
[22] B.K. Lundholt, K.M. Scudder, L. Pagliaro, J. Biomol. Screen. 8 (2003) 566–570.
[23] P. Ramo, R. Sacher, B. Snijder, B. Begemann, L. Pelkmans, Bioinformatics 25 (2009) 3028–3030.
[24] T.R. Jones et al., BMC Bioinformatics 9 (2008) 482.
[25] L. Vincent, P. Soille, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 383–598.
[26] N. Otsu, IEEE Trans. Syst. Man. Cybern. 9 (1979) 62–66.
[27] E.W. Dijkstra, Numer. Math. 1 (1959) 269–271.
[28] A.E. Carpenter et al., Genome Biol. 7 (2006) R100.
[29] S.A. Rifkin, Methods Mol. Biol. 772 (2011) 329–348.
[30] T. Trcek et al., Nat. Protoc. 7 (2012) 408–419.
[31] F. Mueller et al., Nat. Methods 10 (2013) 277–278.
[32] J. Matas, O. Chum, M. Urban, T. Pajdla, Image Vis. Comput. 22 (2002) 761–767.
[33] P. Ruusuvuori et al., BMC Bioinformatics 11 (2010) 248.
[34] E. Bertin, S. Arnouts, Astron. Astrophys. Sup. 117 (1996) 393–404.
[35] S. Mili, K. Moissoglu, I.G. Macara, Nature 453 (2008) 115–119.

7. Control of transcript variability in single mammalian cells.

By

Nico Battich*, Thomas Stoeger* & Lucas Pelkmans.

Under revision in *Cell.*

*Contributed equally.

All experiments described in this chapter were designed, conducted, and analyzed in equal contribution by Nico Battich and Thomas Stoeger. Thomas Stoeger mainly performed the analysis presented in Figure 5, while Nico Battich mainly performed the analysis and experiments presented in Figure 6. The text of this chapter was written in equal contribution by Nico Battich and Thomas Stoeger.

Control of Transcript Variability in Single Mammalian Cells

Nico Battich[1,2*], Thomas Stoeger[1,2*], Lucas Pelkmans[1]

[1]Faculty of Sciences, Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland.

[2]Systems Biology PhD program, Life Science Zurich Graduate School, ETH Zurich and University of Zurich, Zurich, Switzerland.

*These authors contributed equally to this work.

Correspondence to: lucas.pelkmans@imls.uzh.ch

Abstract

A central question in biology is whether variability between genetically identical cells exposed to the same culture conditions is largely stochastic or deterministic. Using image-based transcriptomics in millions of single human cells, we find that while variability of cytoplasmic transcript abundance is large, it is for most genes minimally stochastic, and can be predicted with multivariate models of the phenotypic state and population context of single cells. Computational multiplexing of these predictive signatures across hundreds of genes revealed a complex regulatory system that controls the observed variability of transcript abundance between individual cells. Mathematical modeling and experimental validation show that nuclear retention and transport of transcripts between the nucleus and the cytoplasm is central to buffering stochastic transcriptional fluctuations in mammalian gene expression. Our work indicates that cellular compartmentalization confines transcriptional noise to the nucleus thereby preventing it from interfering with the control of single-cell transcript abundance in the cytoplasm.

Introduction

Gene expression in isogenic cells exposed to the same conditions is heterogeneous, a phenomenon referred to as gene expression noise (Eldar and Elowitz, 2010; Raj and van Oudenaarden, 2008). The origin of this noise can be divided between intrinsic and extrinsic sources (Elowitz et al., 2002; Swain et al., 2002). Intrinsic noise is seen as the inherent consequence of stochastic fluctuations in biochemical reactions and interactions between the components that transcribe and translate genes into mRNA and proteins, respectively (Eldar and Elowitz, 2010; Raj and van Oudenaarden, 2008). For instance, stochastic switching of promoters between a closed, transcription-prohibiting state and an open, permissive, state can lead to bursts in transcription and consequently large variations in transcript abundance between individual cells (Golding et al., 2005; Raj et al., 2006; Suter et al., 2011; Zenklusen et al., 2008). Extrinsic noise is defined as noise that originates from upstream variations in the cellular state that result in higher or lower rates of gene expression or degradation, and is usually the major source of cell-to-cell variability (Altschuler and Wu, 2010; Raser and O'Shea, 2005; Snijder and Pelkmans, 2011). Extrinsic noise is not necessarily of a stochastic nature, but is often considered and modeled stochastically given the complexity of the involved processes, an apparent stochasticity in distributions of single-cell measurements, and an assumed inability to predict these variations at the single-cell level.

Recently, it was shown in human cells that transcript abundance scales with cellular volume (Kempe et al., 2015; Padovan-Merhar et al., 2015), which can be highly variable between single human cells of the same population. Cellular volume is thus an important source of extrinsic noise in gene expression, as has been observed previously in yeast (Newman et al., 2006; Raser and O'Shea, 2004). Similarly, mitochondrial content, which is known to vary between individual mammalian cells, has been identified as a source of extrinsic noise (das Neves et al., 2010). In proliferating mammalian cells that adapt to their multicellular context,

cell-to-cell variability in these and other properties is strongly influenced by the available space to expand cell surface and volume, the relative location of a cell within a population, its local crowdedness, the amount and type of physical force it experiences, the extent by which it faces empty space, and its position in the cell cycle (Dupont et al., 2011; Engler et al., 2006; Frechin et al., 2015; Kafri et al., 2013; Snijder et al., 2009). Since numerous signaling pathways that sense the cellular state and phenotypic properties of single cells and their microenvironment exist, this can result in large-scale adaptation of the transcriptome in single isogenic cells experiencing the same culture conditions. This raises the question to which extent variability in transcript abundance in mammalian cells is of a deterministic nature and can be predicted once the relevant variables of single cells that drive such adaptation are known. Particularly in the context of development and tissue homeostasis, where tight control of gene expression at the single-cell level is required, such variables could influence cell fate decisions that may have previously been considered fully stochastic (Arias and Hayward, 2006; Graf and Stadtfeld, 2008; Macarthur et al., 2009). Furthermore, if most variability in transcript abundance in mammalian cells can be predicted, it raises the question of how stochastic fluctuations that arise during transcription are effectively filtered out while deterministic variability is maintained.

Addressing these questions requires highly accurate measurements of single-cell transcript abundance. A suboptimal efficiency in detecting an individual transcript molecule in a single cell yields for most transcripts single-cell distributions that are largely determined by random detection error (Shapiro et al., 2013). Since single-cell RNA-sequencing has detection efficiencies between 5-20% (Deng et al., 2014; Grun et al., 2014), it cannot be used for sensitive analysis of sources of cell-to-cell variability in transcript abundance. Equally important for obtaining highly accurate measurements for large numbers of single cells is to avoid sampling bias of the cellular states and microenvironments experienced by single cells

in a population (Battich et al., 2013). Furthermore, it is essential to quantify features of the cellular state and microenvironment of the same single cell in which transcript abundance is being measured. Finally, such measurements are ideally obtained for a large number of genes to compare distributions and identify common and gene-specific variables that determine cell-to-cell variability in transcript abundance.

Here, we applied image-based transcriptomics, a high-throughput automated single-molecule fluorescence *in situ* hybridization (sm-FISH) method that we recently developed (Battich et al., 2013), which meets these requirements. Using large-scale single-cell datasets acquired with this approach, we show that cell-to-cell variability in cytoplasmic transcript abundance in human adherent cells can be accurately predicted at the single-cell level with a multivariate set of features that quantify properties of the cellular state and microenvironment, and we experimentally verify some of the underlying causality. We find that for most genes the unexplained variability in cytoplasmic transcript abundance approaches a limit of minimal stochasticity imposed by a Poisson process. The few genes that deviate from this limit also show a high amount of explained variability, suggesting high-level regulation rather than high stochasticity. Through computational multiplexing, we uncover the existence of multi-level transcript homeostasis in single cells to achieve specific adaptation of transcript abundance to the cellular state and microenvironment, according to function of the proteins they encode. Finally, we show that the mammalian nucleus acts as a potent and global buffer of stochastic fluctuations arising from bursts in gene transcription by temporally retaining transcripts in the nucleus. This explains how cytoplasmic transcript abundance in mammalian cells can be minimally stochastic, while deterministic variation is maintained.

Results

Single-cell distributions of cytoplasmic transcript abundance in a human cancer-derived cell line and primary keratinocytes

To study cell-to-cell variability of transcript abundance in human cells, we applied image-based transcriptomics to HeLa cells and freshly isolated primary keratinocytes. This approach uses branched DNA oligonucleotide probes in high-throughput automated single-molecule fluorescence *in situ* hybridization (sm-FISH) (Battich et al., 2013). It provides high-quality images of large numbers of single cells in which each transcript is visible as a bright spot that can be robustly detected, resolved from other spots, and assigned to the corresponding cell using fully automated computer vision algorithms (Battich et al., 2013; Stoeger et al., 2015) (Figure 1A). As a result, accurate and reproducible transcript counts in the cytoplasm of millions of single cells and thousands of genes are obtained (Battich et al., 2013). When visualized across cell populations, this reveals gene-specific, patterns in single cells as shown for *UBE2C*, an ubiquitin-conjugating enzyme that targets cyclins for degradation (Figure 1B, S1A).

In both cell types, we identified, in a semi-automated manner, 5 classes of single-cell distributions of cytoplasmic transcript abundance across the 932 genes. These distributions can be compared to each other and visualized on http://image-based-transcriptomics.org. The vast majority of genes show a unimodal distribution (Figure 1C-D, S1B), which shifts from a one-tailed distribution (class 2) to a skewed two-tailed distribution (class 3) as the mean cytoplasmic transcript abundance increased. This trend occurs despite the theoretical possibility to reach the same mean abundance of transcripts with any of those classes (Munsky et al., 2012). Genes displaying a skewed two-tailed distribution with a broad or flattened peak (class 4) enrich for genes acting during the replication of DNA (8 of 9 genes in HeLa and 5 of 7 genes in keratinocytes). Rarely (1.6% in HeLa and 2.8% in keratinocytes), bimodal distributions were observed, with either one mode representing no expression and

the other mode expression (class 1), or with both modes representing two different levels of expression (class 5). Although we report a lower level of bimodality in the distributions of genes than previously reported for single human cells (Shalek et al., 2013), this likely results from the fact that in our experiments cells are unperturbed and at quasy steady-state, and did not experience a sudden change in culture conditions (e.g. addition of growth factor after serum starvation). Concordantly, the majority of the genes that show bimodal distributions under these culture conditions act during the M phase of the cell cycle (60.0% of HeLa class 5, 71.4% of keratinocytes class 1 and 50.0% of keratinocytes class 5). We also noticed that the coefficient of variation (CV) in single-cell transcript abundance decreased in both cell types monotonically from ~2 to ~0.3 as the mean transcript abundance increased, with only a few outlier genes (3-6%) that show a higher CV than the bulk (Figure 1E). Expectedly, these outliers are enriched in the one-tailed and bimodal distributions of cytoplasmic transcript abundance (class 1, 2, and 5) (Figure 1F).

Cytoplasmic transcript abundance in single human cells can be predicted and is minimally stochastic

In addition to cytoplasmic transcript abundance, we collected from each single cell a multivariate set of 183 features that quantify properties of cell and nucleus shape and area, of protein, DNA and mitochondrial content and texture, and of the extent of local cell crowding, number of neighbors, and relative location to other cells and to empty space in the cell population (Figure 2A). For genes that are expressed (mean transcript abundance per cell > ~4, Figure S2A) we observed that many of these features show a correlation with transcript abundance (Figure S2B), prompting us to investigate the extent by which these features can collectively predict cytoplasmic transcript abundance in single cells. To address this, we learnt data-driven models for each gene on one dataset using multi-linear regression (MLR)

in a principal component (PC)-reduced multidimensional space of the multivariate feature set (Figure 2A). When learning MLR models per gene, increasing numbers of PCs were added until maximum prediction strength of cytoplasmic transcript abundance was reached (Figure 2A). Generally, MLR models consisted of ~20 PCs, which quantify a variety of different aspects of individual cells. For example, the first 6 PCs of HeLa cells consist of features of local cell crowding, distance of cells to each other, their number of neighbors, distance to a cell islet edge, cell and nuclear area, cell volume (as measured by protein content, see Figure S2C), mitochondrial content, DNA content (indicating position in the cell cycle), nuclear morphology, cell shape, and the activity (transcript abundance) of neighboring cells (Figure 2A). In keratinocytes, the first 6 PCs contain somewhat different loadings, such as protein concentration in PC5, but are highly comparable (Figure S2D). Higher PCs used in the MLR models often contain highly specific properties related to cell shape, texture or microenvironment (not shown).

Next, we tested the performance of each MLR model by directly predicting cytoplasmic transcript abundance in each single cell of an independently obtained (~3 weeks later) biological replicate dataset for the same gene, and comparing single-cell predictions with single-cell measurements. The models accurately reproduced single-cell distributions of cytoplasmic transcript abundance (Figure 2B, S2E-F). More importantly, they also achieved high prediction strength (*pS;* coefficient of determination corrected for different impact of technical variability on genes with different transcript abundance, see Extended Computational Procedures) at the single-cell level (Figure 2C-E, S2G). The median *pS* was slightly higher in monoclonal HeLa cells (0.503) than in freshly isolated primary keratinocytes (0.400), possibly due to uncontrolled clonality of the latter cells. Partial least squares regression as well as a non-linear approach using random forests (Liaw and Wiener,

2002) on the non-transformed multivariate feature set achieved virtually identical results (not shown), indicating the robustness of these statistical models.

The *pS* increased as mean cytoplasmic transcript abundance increased, with a median *pS* of 0.29 for low-abundant transcripts (3.7-7.4 mean transcripts per cell) and a median *pS* of 0.71 for high-abundant transcripts (>149 mean transcripts per cell, see Figure S2H). Furthermore, as the examples of *KIF11* (a kinesin involved in spindle formation and chromosome positioning during mitosis), *ERBB2* (a receptor tyrosine kinase that dimerizes with epidermal growth factor receptors), and *CLOCK* (a transcription factor that regulates circadian rhythms) show, the MLR models predict the observed patterns of single-cell expression in cell populations remarkably well, even for low-abundant transcripts (Figure 2D-E, S2G). Naturally, three-state stochastic models of transcription can only reproduce distributions (Figure 2B, S2E-F), and do not have any single-cell prediction strength (Figure 2C-E, S2G) nor can they reproduce single-cell expression patterns in cell populations (Figure 2E, S2F).

Strikingly, when we quantified the amount of variability in cytoplasmic transcript abundance that the MLR models could not explain (see Extended Computational Procedures) (Elowitz et al., 2002), we observed that it approaches a limit of minimal stochasticity as described by a simple one-step Poisson process (Figure 3A, S3A). This was also the case for low-abundance transcripts, and agrees with their lower observed *pS*, since single-cell predictability is more strongly affected by minimal stochasticity when mean levels are low (Figure 3A, S3A). Although some genes did not fall on this limit, we observed that genes whose unexplained variability was furthest away from the Poisson limit, also displayed the highest amount of explained variability (outliers of both increased $\eta^2_{Explained}$ and $\eta^2_{Unexplained}$) (Figure 3B, S3B). This shows that also for these genes, cell-to-cell variability in cytoplasmic transcript abundance originates largely from regulatory processes rather than from intrinsic stochastic sources (Figure 3B).

These results show that cytoplasmic transcript abundance of genes can be accurately predicted at the single-cell level in mammalian adherent cells, both in a cancer-derived laboratory-adapted cell line and in primary cells freshly isolated from a human donor. Single-cell prediction is achieved with features that quantify a variety of different aspects of the phenotypic state of individual cells, their population context and their microenvironment. The amount of cell-to-cell variability that these features cannot predict approaches for most genes a single-step Poisson limit. This suggests that somewhere along the complex life of an RNA molecule, noise buffering occurs to ensure that cytoplasmic transcript abundance becomes minimally stochastic.

Causality between predictors and single-cell transcript abundance

High prediction strength, which indicates a high correlation between predictors and single-cell transcript abundance, does not reveal the presence or direction of causality. For instance, it may be that stochastic fluctuations in transcript abundance of a gene influence the phenotypic state or the population context and microenvironment of an individual cell, such that they become good predictors of these fluctuations. On the other hand, variability in these properties may directly influence the transcript abundance of genes. In growing adherent cell populations, the situation is more complex, involving multiple feedbacks acting at multiple timescales between transcript abundance and cellular phenotype, which emerge as cells proliferate to form populations (Frechin et al., 2015; Snijder and Pelkmans, 2011; Warmflash et al., 2014). To reveal the dominant direction of causality in this situation, we used four orthogonal approaches.

First, we applied Bayesian network inference on the initial datasets (Figure S4A), focusing on four dominant and strong predictors (cell area, protein content/cell volume, DNA content,

and cell crowding). For 83% of the genes where Bayesian networks could reproducibly be inferred, cytoplasmic transcript abundance was placed downstream of one or multiple single-cell features (Figure S4A). The remaining genes (17%) were placed in between, being downstream of cell area or protein content, and upstream of DNA content or cell crowding, which often correlated with gene function. For instance, the cytoplasmic transcript abundance of both *POLA1*, the catalytic subunit of DNA polymerase, and *CDK1*, a cyclin-dependent kinase critical for progression of cells from G1 into S, were placed upstream of DNA content. While cytoplasmic transcript abundance of *ACTR2* and *ACTR3*, the two subunits of the Arp2/3 complex involved in actin polymerization, as well as *RHOA*, a central GTPase in the regulation of the cellular cytoskeleton and adhesion, were placed upstream of local cell crowding (Figure S4B,C). This corresponds to their well-characterized roles in wound healing, cell polarization, collective cell migration, and epithelial-mesenchymal transition, processes that all involve changes in cell shape and crowding (Etienne-Manneville and Hall, 2002) We also quantified bulk nascent transcript synthesis in single cells that were seeded at different numbers per well, inferred a Bayesian network from these measurements, and compared the resulting network with one that is a combination of all single-gene networks. Both networks revealed that cell area and protein content are major determinants of bulk nascent transcript synthesis and cytoplasmic transcript abundance, which are in turn determined by population context effects that arise from the number of cells seeded and DNA content, the latter reflecting position in the cell cycle (Figure S4D).

Second, we grew cells on micropatterns, which constrain the available area for a single cell to spread on, resulting in a strong reduction in the cell-to-cell variability of many single-cell features, particularly in cell size and morphology (Figure 4A,B). Based on this, we used the MLR models to predict which genes would display the strongest reduction in variability in cytoplasmic transcript abundance in cells grown on the smallest micropatterns, and selected

from these 9 genes covering different biological processes (Figure 4A). For all genes, and as exemplified by *RELA*, a subunit of the major transcription factor NF-κB, we observed that constraining the phenotypic state of single cells results in a strong reduction of cell-to-cell variability in cytoplasmic transcript abundance, approaching a Poisson distribution (Figure 4C,D). Strikingly, the small amount of remaining cell-to-cell variability in transcript abundance was accurately predicted based on the small amount of variability remaining between single cells grown on micropatterns (Figure 4C). This shows that constraining single-cell features directly constrains cell-to-cell variability in transcript abundance

Third, we performed systematic RNA interference against 367 genes using 3 independent siRNAs per gene in 6 biological and 3 technical replicates. This did not reveal any relationship between the extent to which two dominant features, nuclear area and cell crowding, correlate with cytoplasmic transcript abundance of a gene and the effect that silencing of this gene had on these two features (Figure S4E). The few genes whose silencing resulted in strong effects were all essential for cell viability, leading to reduced population sizes (Figure 4D), which indirectly changes nuclear area and cell crowding (Snijder et al., 2012).

Fourth, we performed gene induction experiments (Figure 4E). Cells grown for 72 hours to establish heterogeneity in population context and cellular state were serum-starved for 24 hours and subsequently treated with epidermal growth factor (EGF). At 20, 40 and 80 min after induction, we fixed cells and performed image-based transcriptomics on 8 genes induced by EGF. We then learnt MLR models on each time-point after induction as well as on the serum-starved non-induced state, and used these to predict single-cell cytoplasmic transcript abundance in a replicate experiment (Figure 4E). While *pS* was lower in the non-induced state or in the presence of serum, it was higher at peak expression level following induction, matching the global trend that *pS* scales with transcript abundance (Figure 4F,

S4F). As shown for *JUN* and *FOS*, two immediate early response genes (Morgan and Curran, 1995), the MLR models accurately reproduced the change in distributions of cytoplasmic transcript abundance during induction, including the emergence of bimodality, as well as single-cell patterns of EGF-induced gene expression in cell populations (Figure 4G-H and S4G). Strikingly, MLR models learnt on serum-starved non-induced cells were able to predict the single-cell expression patterns in induced cell populations, when correcting for difference in mean expression levels (Figure 4G-H). This shows that it is largely the predetermined phenotypic state or microenvironment of a single cell that determines its response to EGF.

Together, these experiments and analyses show that in human adherent cells grown in culture, the emergence of heterogeneity in phenotypic state, population context, and microenvironment of single cells is the dominant source of cell-to-cell variability in cytoplasmic transcript abundance, making it for most genes largely predictable. This does not only apply to cells at quasi steady-state continuously grown in serum, but also, and more profoundly, during an acute induction of gene expression by EGF, also when this leads to bimodal gene expression.

Computational multiplexing of cytoplasmic transcript abundance reveals multi-level transcript homeostasis in single cells.

We next studied the biological information that the MLR models contain, taking advantage of their generally high prediction strength at the single-cell level. This allowed us to perform computational multiplexing (Figure 5A), in which we predicted the transcript abundance of one gene in each cell of a population in which we had measured the transcript abundance of another gene. In this manner, we could calculate pairwise correlations between the predicted

and measured single-cell transcript abundances for ~$2.5x10^5$ gene-gene combinations across ~5,000 single cells (Figure 5A, S5A). We then calculated the similarity between two genes in their pairwise single-cell correlations with all other genes, and created a similarity matrix for each cell type. The matrices contained a high degree of modality with various sub-clusters (Figure 5A, Supplemental Table 1), presenting a systems-level map of single-cell transcript homeostasis in human adherent cell populations (Figure S5B, C).

To visualize this map, we created a gene interaction network for each cell type, in which two genes are connected when they were within the top 2% highest similarity scores (Figure 5A, B, S5A). To reveal patterns in these networks, we first looked at the two most dominant predictors, cell area and cell volume. Plotting the ratio of the correlation of cytoplasmic transcript abundance with these two predictors on the networks revealed areas of genes with a higher correlation to cell volume, or higher correlation to cell area (Figure 5B, S5D, Supplemental Data 1). The latter are strongly enriched in genes encoding for proteins that contain a signal peptide, a transmembrane domain, or that are N-glycosylated (Figure 5B), as well as for cytosolic proteins with important membrane-related functions (not shown). This indicates the existence of mechanisms that allow distinct adaptation of transcript abundance to the volume or surface area of a single cell, depending on whether it encodes for a protein with cytosolic or membrane-related functions.

We next plotted on the networks the mean absolute correlations of transcript abundance to selected sets of features related to the population context, to cell size and shape, DNA content, and neighbor activity (Figure 5C). This revealed multiple sub-regions in the networks that consist of groups of genes whose cytoplasmic transcript abundance is adapted in specific ways to different combinations of features. For example, a particularly outstanding sub-cluster present in both networks (K1 in keratinocytes and H1 in HeLa, see Figure 5D,

S5E), shows high correlations with features of DNA content and texture and nuclear morphology, and is enriched in genes that function in the cell cycle.

We also noticed that within dense regions of the networks, highly differentiated and specific adaptation is visible. For instance, sub-cluster K2 and K3 lie next to each other in the keratinocyte network (Figure 5C). Sub-cluster K2 consists of genes whose transcript abundance shows a specific and strong correlation with neighbor activity, and contains immediate early genes (e.g. *JUN*) including secreted molecules (e.g. *VEGFA* and *DKK1*) (Figure 5D). A similar sub-cluster was also found in HeLa cells (H2, see Figure S5E). H2 contains genes that display high levels of both explained and unexplained variability (compare to Figure 3B), including the early response genes in the EGF induction experiment (such as *JUN*, *FOS*, and *NR4A2*). This indicates that both cell types show a highly variable expression of a group of genes that respond quickly to signals in a correlated manner determined by the activity of cell neighbors, suggesting the involvement of paracrine signaling (Avraham and Yarden, 2011). Sub-cluster K3 consists of genes whose single-cell transcript abundance shows strong correlation with multiple sets of selected features, including those of the population context, of cell size and shape, of mitochondrial abundance, nuclear morphology and also neighbor activity (Figure 5D). It contains 10 of the 13 mitochondrially-encoded protein-coding genes, indicating that multi-level control of single-cell transcript abundance also occurs for genes not transcribed in the nucleus.

The high degree of modularity and the presence of multiple subgroups of genes whose transcript abundance is adapted in highly differentiated and specific ways in single cells exposed to the same conditions, demonstrates the existence of a complex multi-level transcript homeostasis system that drives cell-to-cell variability in gene expression. This ensures that levels of transcripts are precisely adapted to the physiological state of a single

cell and its microenvironment according to the function of the RNA or the protein they encode for.

Transcript retention in the nucleus and export into the cytoplasm efficiently buffers stochastic bursts in gene transcription

On first sight, the high degree of predictability in cytoplasmic transcript abundance at the single-cell level contradicts the view that it arises from stochasticity in gene transcription, caused by, amongst others, the stochastic switching of promoters between a closed transcription-prohibiting state and an open permissive state. In reconciling our findings with this view of transcription at the single-cell level (Raj and van Oudenaarden, 2008), we realized that bDNA sm-FISH, unlike most methods that quantify single-cell transcript abundance, specifically detects transcripts in the cytoplasm. This raises the possibility that random fluctuations in transcript abundance arising from bursts in transcription are filtered out during nuclear processing and/or export from the nucleus to the cytoplasm (Singh and Bokes, 2012; Xiong et al., 2010). In electronics and telecommunication, it is well known that compartmentalization, as for instance used in the leaky bucket algorithm, provides an efficient and general mechanism to eliminate stochastic burst-like noise (jitter) in signals (Tanenbaum, 2003). This requires that the rate of signal output is relatively slow and constant, and the compartment has considerable storage capacity to act as a buffer for stochastically fluctuating input. Such requirements may also be fulfilled by the mammalian nucleus, which is relatively large, contains a high concentration of RNA molecules (Piwnicka et al., 1983), retains nascent RNA transcripts for further processing, and has a highly constant and relatively low density of nuclear pores (Maul et al., 1972). Furthermore, the few measurements that exist on nuclear export dynamics of individual transcripts suggest that, at a given moment in time, a single RNA molecule has a low probability of being exported

(Grunwald and Singer, 2010). In addition, nuclear retention has been described as a mechanism of regulating gene expression (Prasanth et al., 2005).

To test if nuclear compartmentalization can theoretically act as a noise buffer in mammalian cells, we developed an agent-based single-cell mathematical model and performed computer simulations inspired by the leaky bucket algorithm (Figure 6A). In the model, gene activation and transcription is governed by a three-state stochastic model (Neuert et al., 2013; Raj and van Oudenaarden, 2008), in which the gene switches randomly between an 'off' state (S3) and a transcription-competent state (S2), which switches randomly to a transcription-initiated (S1) or 'on' state and back. Once transcription is initiated, RNA synthesis occurs at randomly fluctuating transcription rates. The time spent in the 'on' and 'off' states can be varied. Each transcript is then retained for a certain amount of time in the nucleus, after which it is transported into the cytoplasm. Nuclear retention time is used as a general term to comprise the various events between birth of a single transcript molecule and its emergence into the cytoplasm, including chromatin dissociation, nuclear diffusion, processing, and binding to and transport across the nuclear pore. It is modeled as a combination of a 3D diffusion process and a probabilistic interaction with the nuclear pore, and can be varied. Nuclear degradation of transcripts is not considered. Finally, transcript degradation in the cytoplasm is modeled as a single probabilistic function that can also be varied (Figure 6A). To quantify the effect that nuclear retention has on the amount of stochasticity in transcript abundance in this model, time distributions between simulated transcript production events (dT_s) and between nuclear export events (dT_e) are obtained, and the distance of these distributions to a Poisson distribution determined (Figure 6B). Physiological boundaries for nuclear retention times of transcripts were obtained from a recently collected high-quality RNA-seq dataset on LPS-induced transcription in mouse bone marrow-derived macrophages (Bhatt et al., 2012). From 282 genes, we derived a nuclear retention time of newly synthesized transcripts

between ~5-90 min, with a median of ~20 min (Figure 6C). Because these genes are enriched in fast-responding genes during stress signaling in macrophages, these nuclear retention times are an underestimation for most other genes.

We performed model simulations with different burst-like gene transcription scenarios, ranging from transcription 'on' times (corresponding to state S1 in the model of the gene module, Figure 6A) between 5 - 20 min and 'off' times (corresponding to state S2 or S3 in Figure 6A) between 5 - 60 min, which lie in the range of observed bursting dynamics of endogenous genes in mammalian cells (Ochiai et al., 2014). Furthermore, longer 'on' times do not reflect burst-like gene expression and are already close to a Poisson limit during synthesis in the nucleus (Figure 6D). Longer 'off' times, in our opinion, reflect non-stochastic regulation such as refractory periods, feedbacks, or oscillating cellular states (Sanchez and Golding, 2013), which nuclear retention should not filter out. In this range, variability in transcript synthesis is far away from the Poisson limit (Figure 6D). However, export of the produced transcripts into the cytoplasm was efficiently converted into a Poisson process as mean retention time increased (Figure 6D, S6A). Over all bursting scenarios, a mean nuclear retention time of 15 min was able to buffer ~57% of the stochastic fluctuations introduced by bursts, which increased to ~90% at 40 min of mean nuclear retention time. Importantly, when we modeled bursting scenarios with 'on' and 'off' times of ~5.5 min (scenario 1 in Figure 6D), we observed ~50% buffering already at a mean nuclear retention time of 6 min. This corresponds to the measured induction and nuclear retention times of *FOS* and *JUN*, which are between 6 and 10 min (Figure 6C), indicating that the short retention times observed for fast-responding genes could also have a noise buffering effect. This theoretical analysis indicates that for most genes, nuclear retention is long enough to reduce stochastic variation arising from bursts in transcription, also for immediate early genes. Furthermore, it suggests that molecular mechanisms may have evolved that gene

specifically couple the time scales of transcript retention in the nucleus to the rate of their induction (Culjkovic et al., 2006). This would allow fast response times for early immediate genes, but still ensure non-stochastic regulation of transcript abundance in the cytoplasm.

To test experimentally whether nuclear retention increases the predictability of cytoplasmic transcript abundance in single mammalian cells, we adapted bDNA sm-FISH to detect transcripts in the nucleus, and performed an EGF induction experiment where we measured both nuclear and cytoplasmic transcript abundance. As expected, the increase in cytoplasmic transcript abundance of genes reacting to EGF followed with a delay the increase in nuclear transcript abundance (Figure S6B-C). Moreover, bursts of transcription were clearly visible in the nucleus (Figure S6C). Importantly, we found that the coefficient of variation (CV^2) was higher in the nucleus than in the cytoplasm, in particular in cases when transcripts were less abundant. This decrease in transcript variability in the cytoplasm compared to the nucleus was predicted by the model (Figure 6E). Moreover, we found that MLR models have higher prediction strength on cytoplasmic transcript abundance than on nuclear transcript abundance, (e.g. 2.5-fold higher for *FOS* and *JUN*), which is consistent with the agent-based model (Figure 6F).

We next used long-term time-lapse imaging of single HeLa cells expressing an inducible transcript containing 24 bacteriophage MS2 stem loops, as well as Halo-tagged MS2 coat protein, which binds to the stem loops (Halstead et al., 2015). Time-lapse imaging carried out for 5-13 hours after gene induction showed repeated bursts of transcription in the nucleus (Figure 6G). We also observed a transient accumulation of transcripts at the inner nuclear envelope, and a delay between the increase in nuclear transcript abundance and cytoplasmic transcript abundance, both indicative of nuclear retention (Figure 6G). From the movies, we estimated that the length of bursts (‘on’ time) were 10-60 min, the intervals between bursts (‘off’ time) were 20-100 min, and nuclear retention time was ~40 min. These values are all

within the range of the modeled parameter space, indicating that cytoplasmic transcript abundance should display less stochastic variability than nuclear transcript abundance. To measure this within the same single cells, we calculated the auto-correlation in transcript abundance over time in both the nucleus and the cytoplasm. A low auto-correlation is indicative of stochastic fluctuations. Consistent with the model's predictions, we observed that auto-correlation measurements of transcript abundance over up to 1-hour time periods are higher in the cytoplasm than in the nucleus (Figure 6H, S6D). This directly shows that during gene induction, transcript abundance shows more stochastic fluctuations over time in the nucleus than in the cytoplasm, indicative of buffering through nuclear compartmentalization and retention.

Taken together, the combined theoretical and experimental approach showed that cellular compartmentalization separating the nucleus from the cytoplasm is an efficient mechanism to dampen stochastic fluctuations arising from bursts in gene transcription for most genes. This explains how cytoplasmic transcript abundance in single cells can approach a Poisson limit of minimal stochasticity despite the occurrence of burst-like gene transcription.

Discussion

In this study, we perform highly accurate measurements of transcript abundance in large numbers of single adherent human cells with single-molecule resolution for a thousand genes using image-based transcriptomics. We combine these measurements with a multivariate set of features from the same single cells that quantify multiple properties of the cellular state, their population context, and their microenvironment. We show that multi-linear regression models based on these features can predict single-cell distributions, have high prediction strength on single-cell transcript abundance, and can accurately predict single-cell expression

patterns. This is in contrast with stochastic models of gene transcription, which can only reproduce single-cell distributions. We show that the amount of variability not explained by multi-linear regression approaches a system of minimal stochasticity given by a Poisson process. We reveal that the causality underlying this high predictability stems from mechanisms by which the cellular state, the population context, and the microenvironment determine cytoplasmic transcript abundance in single cells, for which we provide a systems-level map across several hundred genes. Finally, we show that mammalian cells can achieve minimal stochasticity in cytoplasmic transcript abundance by means of nuclear compartmentalization, which, through temporally retaining transcripts in the nucleus, provides a general and potent mechanism to buffer stochastic fluctuations caused by bursts in gene transcription.

Our findings pertain to virtually all of the genes analyzed in adherent human cells, both when cells are at quasi steady state in the continuous presence of serum, as well as during acute gene induction experiments after a period of serum starvation. This illustrates that even at time-scales of less than 1 hour, a differential response in the up-regulation of cytoplasmic transcript abundance in single adherent mammalian cells is largely of non-stochastic origin. Only a few genes display simultaneously a high degree of explainable variability as well as a high degree of unexplainable variability. These are immediate early response genes, the transcripts of which accumulate rapidly in the cytoplasm after induction of expression, are only shortly retained in the nucleus and are subject to high cytoplasmic turnover. While this limits the nucleus' ability to completely filter out stochastic variability caused by bursts in gene transcription for these genes, their relatively brief nuclear retention still has a sufficient noise dampening effect. Thus, while cell-to-cell variability in cytoplasmic transcript abundance in mammalian cells is often large, our findings show that the cause of this

variability is not stochastic, but is determined by a multi-level system regulating transcript homeostasis in single cells.

The use of nuclear retention for noise filtering underscores the notion that mammalian cells do not rely on the induction of gene transcription for very fast responses (Alberts, 2008). For the fastest responding genes in mammalian cells, such as *FOS* and *JUN*, nuclear retention times appear adjusted to the rate of induction, short enough to minimize the delay in response, but long enough to enable efficient noise buffering.

In prokaryotes, where a nucleus is absent and RNA pre-processing is minimal, transcriptional responses can make use of co-transcriptional translation and can thus be very fast (Martin and Koonin, 2006). Also in single-cell eukaryotes such as yeast, which show less extensive nuclear processing of transcripts and have considerably smaller nuclei, transcriptional responses may overall be somewhat faster than in mammalian cells (Kresnowati et al., 2006). This suggests that as cells acquired a nucleus during evolution and formed multicellular organisms, the increased complexity in nuclear RNA processing came with the additional benefit of filtering out stochasticity in gene transcription, at a slight expense of response time.

Several mechanisms of buffering noise in mammalian gene expression have been proposed, mostly involving gene-specific solutions such as feedback or feedforward motifs in their transcriptional regulation, or the co-expression of its own microRNA (Arias and Hayward, 2006; Li et al., 2009; Milo et al., 2002; Schmiedel et al., 2015). Cellular compartmentalization into the nucleus and the cytoplasm however acts more globally. It thus seems likely that regulation of nuclear retention is a primary mechanism for noise buffering of gene transcription in mammalian cells, with additional mechanisms, such as incoherent feedforward loops based on microRNAs (Schmiedel et al., 2015), allowing further gene-specific adaptation of variability and preventing stochastic fluctuations to propagate into protein translation. Furthermore, while a relatively slow rate of transcript degradation in

mammalian cells may also contribute to buffering stochastic fluctuations, this will also affect the mean abundance of a transcript as well as the ability to quickly change relative concentrations of a transcript. Buffering through nuclear retention does not or to a much lesser extent have these drawbacks.

As suggested by the broad range of nuclear retention times for individual genes in mammalian cells, a highly adaptive system of fine-tuning nuclear retention time to transcription dynamics may be in place, which is undoubtedly more sophisticated than our simplified leaky bucket model currently assumes. One may envision mechanisms where transcript release from the nucleus, or transcript storage within sub-compartments of the nucleus, is additionally regulated (Bhatt et al., 2012; Culjkovic-Kraljacic et al., 2012; Prasanth et al., 2005; Taddei et al., 2006). For instance, releasing pre-stored nuclear transcripts into the cytoplasm upon a stimulus without the need for transcription may achieve a highly regulated fast response devoid of stochastic burst-like fluctuations. Regulation of the association of active transcription sites to nuclear pores (Taddei et al., 2006), and the direct involvement of nuclear pore components in the regulation of transcription machinery (Schneider et al., 2015) may also play a role in this. A transcriptome-wide comparison of the spatial and temporal dynamics of bursts in gene transcription, rates of transcript synthesis and chromatin release, nuclear retention times, and cytoplasmic turnovers by time-lapse imaging will likely reveal additional mechanisms.

Besides the generally accepted view that nuclear compartmentalization of the genome during the course of evolution allowed more complex gene regulation and the rise of multicellular organisms, we speculate that it provided another important advantage: It allows a buffering of transcriptional noise, resulting in a tighter control of gene expression variability that is essential for successful multicellular development.

Experimental procedures

More details of all experimental and computational procedures are described in the Extended Experimental and Computational Procedures. CellProfiler modules are available on http://github.com/pelkmanslab. Single-cell distributions can be browsed online at http://image-based-transcriptomics.org.

Cell Cultivation

HeLa cells were cultivated and seeded for experiments as described before (Battich et al., 2013). The HeLa cells are a single-cell clone isolated from the HeLa "Kyoto" strain, which has been kindly provided by J. Ellenberg (EMBL, Heidelberg). Keratinocytes were donated by a healthy 2.5-year old male, isolated (Biedermann et al., 2010) and kindly provided by E. Reichmann and L. Pontiggia (UZH, Zurich). Keratinocytes were cultivated in CnT-57 medium (CELLnTEC) supplemented at 1:100 [v:v] with Pen Strep (Gibco). For propagation, but not for image-based transcriptomics (which was performed in multi-well plates), plastic dishes were coated with rat tail collagen I (BD Biosciences). For image-based transcriptomics 1800 keratinocytes were seeded per well and cultivated for 3 days.

Image-based Transcriptomics

Image-based transcriptomics, including sample processing and computational object detection, was performed as described earlier (Battich et al., 2013; Stoeger et al., 2015) except that for keratinocytes a final dilution of protease of 1:2000 was used. Briefly, cells were seeded in 384-well plates and transcripts of distinct genes were stained in separate wells by branched DNA single-molecule fluorescence in-situ hybridization using ViewRNA reagents (Affymetrix) on an automated experimental platform and imaged using a

CellVoyager 7000 (Yokogawa) with an enhanced CSU-X1 spinning disk (Microlens-enhanced dual Nipkow disk confocal scanner, wide view type) and a 40× Olympus objective of 0.95 NA and Neo sCMOS cameras (Andor, 2,560 × 2,160 pixels).

Predictions of Spots per Cell

Multi-linear regression (MLR) of spots per cell were trained using the robustfit function of MATLAB and applied to an independent biological replicate. Stochastic simulations were carried out using the Gillespie algorithm. See Extended Computational Procedures for a detailed description.

In Vivo Imaging of Transcripts

HeLa 11ht cells (Weidenfeld et al., 2009) stably expressing a doxycycline-inducible Renilla luciferase transcript that contains a chimeric β-globin / IgG intron in the 5′ UTR and 24 copies of the MS2 stem-loops in the 3′UTR (HeLa 11ht MS2) were kindly provided by Jeffrey Chao (Friedrich Mischer Institute, Basel). HeLa 11ht MS2 also expressed a NLS-HA-MCP-Halo tag that bound the MS2 stem-loops for transcript detection. Cells were cultivated as described before, but supplemented with 10% doxycycline free FBS. Prior to imaging HeLa 11ht MS2 cells were incubated for 20 min at 37ºC in complete medium with 0.1µM JF549 Halo dye (Grimm et al., 2015) and then induced with 0.1µg/ml of doxycycline in imaging medium (Phenol red free DMEM, Invitrogen, supplemented with 10% FCS, PenStrep and 100µM Trolox). To visualize transcripts tagged with the MS2 stem loops, HeLa 11ht MS2 cells were imaged in an inverted Nikon Eclipse Ti-E microscope equipped with the Yokogawa Spinning Disc System W1 and a Nikon CFI PlanApo 100x oil immersion objective. Cells we imaged using a 561nm laser line and a BP 609/54 emission filter, for 5-

hours or 13-hours, 1-4 hours after induction with a time resolution of 10 min. Six z-planes spaced by 1.2 µm were acquired per time point, to sample MS2 transcripts in full height of the cells. The segmentation of cells and nuclei was curated manually using CellProfiler^{MT}. Spot detection in the cytoplasm was done as described before. To avoid problem in the spot detection of transcripts in the nucleus due to different background in different cells and z-planes, they were counted manually using the aid of customized software written in MATLAB.

Supplemental information

Supplemental Information includes Extended Experimental and Computational Procedures, six figures, and one movie and can be found with this article online.

Acknowledgments

We thank Y. Yakimovich for help with computational infrastructure, R. Holtackers for help with experiments, J. Patterson for assistance, J. Wilbertz (Friedrich Miescher Institute) and J. Chao (Friedrich Miescher Institute), J. Ellenberg (European Molecular Biology Laboratory), E. Reichmann (University of Zurich) and L. Pontiggia (University of Zurich) for reagents, and all members of the lab for useful comments on the manuscript. L.P. acknowledges financial support for this project from SystemsX.ch, the Swiss National Science Foundation, the University of Zurich, and the University of Zurich Research Priority Program in Systems Biology and Functional Genomics.

Contributions

L.P. initiated the study. N.B., T.S. and L.P. designed and analyzed the experiments and wrote the manuscript. N.B. and T.S. performed the experiments. The order of appearance of the first authors of this and a related study (Battich et al., 2013) reflects the outcome of a random event (coin toss).

References

Alberts, B. (2008). Molecular biology of the cell, 5th edn (New York: Garland Science).

Altschuler, S.J., and Wu, L.F. (2010). Cellular heterogeneity: do differences make a difference? Cell *141*, 559-563.

Arias, A.M., and Hayward, P. (2006). Filtering transcriptional noise during development: concepts and mechanisms. Nature reviews Genetics *7*, 34-44.

Avraham, R., and Yarden, Y. (2011). Feedback regulation of EGFR signalling: decision making by early and delayed loops. Nat Rev Mol Cell Biol *12*, 104-117.

Battich, N., Stoeger, T., and Pelkmans, L. (2013). Image-based transcriptomics in thousands of single human cells at single-molecule resolution. Nature methods *10*, 1127-1133.

Bhatt, D.M., Pandya-Jones, A., Tong, A.J., Barozzi, I., Lissner, M.M., Natoli, G., Black, D.L., and Smale, S.T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell *150*, 279-290.

Biedermann, T., Pontiggia, L., Bottcher-Haberzeth, S., Tharakan, S., Braziulis, E., Schiestl, C., Meuli, M., and Reichmann, E. (2010). Human Eccrine Sweat Gland Cells Can Reconstitute a Stratified Epidermis. Journal of Investigative Dermatology *130*, 1996-2009.

Culjkovic-Kraljacic, B., Baguet, A., Volpon, L., Amri, A., and Borden, K.L. (2012). The oncogene eIF4E reprograms the nuclear pore complex to promote mRNA export and oncogenic transformation. Cell reports *2*, 207-215.

Culjkovic, B., Topisirovic, I., Skrabanek, L., Ruiz-Gutierrez, M., and Borden, K.L. (2006). eIF4E is a central node of an RNA regulon that governs cellular proliferation. J Cell Biol *175*, 415-426.

das Neves, R.P., Jones, N.S., Andreu, L., Gupta, R., Enver, T., and Iborra, F.J. (2010). Connecting variability in global transcription rate to mitochondrial variability. PLoS biology *8*, e1000560.

Deng, Q., Ramskold, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science *343*, 193-196.

Dupont, S., Morsut, L., Aragona, M., Enzo, E., Giulitti, S., Cordenonsi, M., Zanconato, F., Le Digabel, J., Forcato, M., Bicciato, S., *et al.* (2011). Role of YAP/TAZ in mechanotransduction. Nature *474*, 179-183.

Eldar, A., and Elowitz, M.B. (2010). Functional roles for noise in genetic circuits. Nature *467*, 167-173.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183-1186.

Engler, A.J., Sen, S., Sweeney, H.L., and Discher, D.E. (2006). Matrix elasticity directs stem cell lineage specification. Cell *126*, 677-689.

Etienne-Manneville, S., and Hall, A. (2002). Rho GTPases in cell biology. Nature *420*, 629-635.

Frechin, M., Stoeger, T., Daetwyler, S., Gehin, C., Battich, N., Damm, E.M., Stergiou, L., Riezman, H., and Pelkmans, L. (2015). Cell-intrinsic adaptation of lipid composition to local crowding drives social behaviour. Nature.

Golding, I., Paulsson, J., Zawilski, S.M., and Cox, E.C. (2005). Real-time kinetics of gene activity in individual bacteria. Cell *123*, 1025-1036.

Graf, T., and Stadtfeld, M. (2008). Heterogeneity of embryonic and adult stem cells. Cell Stem Cell *3*, 480-483.

Grimm, J.B., English, B.P., Chen, J., Slaughter, J.P., Zhang, Z., Revyakin, A., Patel, R., Macklin, J.J., Normanno, D., Singer, R.H., *et al.* (2015). A general method to improve fluorophores for live-cell and single-molecule microscopy. Nature methods *12*, 244-250, 243 p following 250.

Grun, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nature methods *11*, 637-640.

Grunwald, D., and Singer, R.H. (2010). In vivo imaging of labelled endogenous beta-actin mRNA during nucleocytoplasmic transport. Nature *467*, 604-607.

Halstead, J.M., Lionnet, T., Wilbertz, J.H., Wippich, F., Ephrussi, A., Singer, R.H., and Chao, J.A. (2015). Translation. An RNA biosensor for imaging the first round of translation from single cells to living animals. Science *347*, 1367-1671.
Kafri, R., Levy, J., Ginzberg, M.B., Oh, S., Lahav, G., and Kirschner, M.W. (2013). Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle. Nature *494*, 480-483.
Kempe, H., Schwabe, A., Cremazy, F., Verschure, P.J., and Bruggeman, F.J. (2015). The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise. Molecular biology of the cell *26*, 797-804.
Kresnowati, M.T., van Winden, W.A., Almering, M.J., ten Pierick, A., Ras, C., Knijnenburg, T.A., Daran-Lapujade, P., Pronk, J.T., Heijnen, J.J., and Daran, J.M. (2006). When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation. Molecular systems biology *2*, 49.
Li, X., Cassidy, J.J., Reinke, C.A., Fischboeck, S., and Carthew, R.W. (2009). A microRNA imparts robustness against environmental fluctuation during development. Cell *137*, 273-282.
Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. R News *2*, 18-22.
Macarthur, B.D., Ma'ayan, A., and Lemischka, I.R. (2009). Systems biology of stem cell fate and cellular reprogramming. Nat Rev Mol Cell Biol *10*, 672-681.
Martin, W., and Koonin, E.V. (2006). Introns and the origin of nucleus-cytosol compartmentalization. Nature *440*, 41-45.
Maul, G.G., Maul, H.M., Scogna, J.E., Lieberman, M.W., Stein, G.S., Hsu, B.Y., and Borun, T.W. (1972). Time sequence of nuclear pore formation in phytohemagglutinin-stimulated lymphocytes and in HeLa cells during the cell cycle. J Cell Biol *55*, 433-447.
Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science *298*, 824-827.
Morgan, J.I., and Curran, T. (1995). Immediate-early genes: ten years on. Trends Neurosci *18*, 66-67.
Munsky, B., Neuert, G., and van Oudenaarden, A. (2012). Using gene expression noise to understand gene regulation. Science *336*, 183-187.
Neuert, G., Munsky, B., Tan, R.Z., Teytelman, L., Khammash, M., and van Oudenaarden, A. (2013). Systematic identification of signal-activated stochastic gene regulation. Science *339*, 584-587.
Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441*, 840-846.
Ochiai, H., Sugawara, T., Sakuma, T., and Yamamoto, T. (2014). Stochastic promoter activation affects Nanog expression variability in mouse embryonic stem cells. Sci Rep *4*, 7125.
Padovan-Merhar, O., Nair, G.P., Biaesch, A.G., Mayer, A., Scarfone, S., Foley, S.W., Wu, A.R., Churchman, L.S., Singh, A., and Raj, A. (2015). Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms. Molecular cell *58*, 339-352.
Piwnicka, M., Darzynkiewicz, Z., and Melamed, M.R. (1983). RNA and DNA content of isolated cell nuclei measured by multiparameter flow cytometry. Cytometry *3*, 269-275.
Prasanth, K.V., Prasanth, S.G., Xuan, Z., Hearn, S., Freier, S.M., Bennett, C.F., Zhang, M.Q., and Spector, D.L. (2005). Regulating gene expression through RNA nuclear retention. Cell *123*, 249-263.
Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. PLoS biology *4*, e309.
Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216-226.
Raser, J.M., and O'Shea, E.K. (2004). Control of stochasticity in eukaryotic gene expression. Science *304*, 1811-1814.
Raser, J.M., and O'Shea, E.K. (2005). Noise in gene expression: origins, consequences, and control. Science *309*, 2010-2013.
Sanchez, A., and Golding, I. (2013). Genetic determinants and cellular constraints in noisy gene expression. Science *342*, 1188-1193.
Schmiedel, J.M., Klemm, S.L., Zheng, Y., Sahay, A., Bluthgen, N., Marks, D.S., and van Oudenaarden, A. (2015). Gene expression. MicroRNA control of protein expression noise. Science *348*, 128-132.
Schneider, M., Hellerschmied, D., Schubert, T., Amlacher, S., Vinayachandran, V., Reja, R., Pugh, B., Clausen, T., and Köhler, A. (2015). The Nuclear Pore-Associated TREX-2 Complex Employs Mediator to Regulate Gene Expression. Cell *162*, 1016-1028.
Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., *et al.* (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236-240.
Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature reviews Genetics *14*, 618-630.

Singh, A., and Bokes, P. (2012). Consequences of mRNA transport on stochastic variability in protein levels. Biophys J *103*, 1087-1096.
Snijder, B., and Pelkmans, L. (2011). Origins of regulated cell-to-cell variability. Nat Rev Mol Cell Biol *12*, 119-125.
Snijder, B., Sacher, R., Ramo, P., Damm, E.M., Liberali, P., and Pelkmans, L. (2009). Population context determines cell-to-cell variability in endocytosis and virus infection. Nature *461*, 520-523.
Snijder, B., Sacher, R., Ramo, P., Liberali, P., Mench, K., Wolfrum, N., Burleigh, L., Scott, C.C., Verheije, M.H., Mercer, J., *et al.* (2012). Single-cell analysis of population context advances RNAi screening at multiple levels. Molecular systems biology *8*, 579.
Stoeger, T., Battich, N., Herrmann, M.D., Yakimovich, Y., and Pelkmans, L. (2015). Computer vision for image-based transcriptomics. Methods.
Suter, D.M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., and Naef, F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. Science *332*, 472-474.
Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. Proceedings of the National Academy of Sciences of the United States of America *99*, 12795-12800.
Taddei, A., Van Houwe, G., Hediger, F., Kalck, V., Cubizolles, F., Schober, H., and Gasser, S.M. (2006). Nuclear pore association confers optimal expression levels for an inducible yeast gene. Nature *441*, 774-778.
Tanenbaum, A.S. (2003). Computer networks, 4th edn (Upper Saddle River, NJ: Prentice Hall PTR).
Warmflash, A., Sorre, B., Etoc, F., Siggia, E.D., and Brivanlou, A.H. (2014). A method to recapitulate early embryonic spatial patterning in human embryonic stem cells. Nature methods *11*, 847-854.
Weidenfeld, I., Gossen, M., Low, R., Kentner, D., Berger, S., Gorlich, D., Bartsch, D., Bujard, H., and Schonig, K. (2009). Inducible expression of coding and inhibitory RNAs from retargetable genomic loci. Nucleic Acids Res *37*, e50.
Xiong, L.P., Ma, Y.Q., and Tang, L.H. (2010). Attenuation of transcriptional bursting in mRNA transport. Phys Biol *7*, 016005.
Zenklusen, D., Larson, D.R., and Singer, R.H. (2008). Single-RNA counting reveals alternative modes of gene expression in yeast. Nature structural & molecular biology *15*, 1263-1271.

Figure legends

Figure 1. Image-based transcriptomics of cell-to-cell variability in cytoplasmic transcript abundance. (**A**) Scheme of *in situ* detection of single transcript molecules (spots per cell). (**B**) Left side: A HeLa cell population stained for cytoplasmic UBE2C transcripts (bDNA sm-FISH in green). Right side: Visualization of the quantified cytoplasmic transcript abundance (spots per cell) by pseudo-coloring single-cell segmentations. Dashed boxes mark enlargements. Cells are discarded by machine learning (SVM, grey) when they touch image borders or are wrongly segmented. (**C**) Classification of single-cell distributions of cytoplasmic transcript abundance in HeLa cells. Genes are binned by their mean spot number per cell (the mean spot number of each bin is indicated on top). Hatched pattern indicates occurrence of class 2 and class 3 in different subsamples of the observed distributions. (**D**) Gene examples with single-cell cytoplasmic transcript abundance distributions belonging to

the various classes. For more examples see http://image-based-transcriptomics.org. (**E**) Coefficient of variation per gene (dot) as a function of cytoplasmic transcript abundance (mean spots per cell), colored according to their distribution class as in (B). Dashed line defines outliers exceeding one standard deviation of a LOESS fit. (**F**) Enrichment for cytoplasmic transcript abundance distribution classes among outlier genes over non-outlier genes. Asterisks indicate Fisher's exact test below 0.05. See also Figure S1.

Figure 2. Predicting cytoplasmic transcript abundance in single cells within a population. (**A**) Overview of extracted features describing the cellular state, population context and microenvironment of single cells (right), the loadings of the first 6 principal components of this multivariate feature space (middle), the construction of multi-linear regression (MLR) models using principal components, and the calculation of prediction strength (*pS*), taking into account technical noise. (**B**) Prediction of single-cell transcript distributions of *KIF11*, *ERBB2* and *CLOCK* in HeLa cells by MLR models and three-state stochastic models. (**C**) Distribution of prediction strengths (*pS*) for 583-598 genes using MLR models (black filled bars) and three-state stochastic models (dashed open bars). Size of each bin is 0.1. (**D**) Prediction of *KIF11*, *ERBB2*, and *CLOCK* cytoplasmic transcript abundance in single HeLa cells by MLR models and three-state stochastic models. (**E**) Visualization of measured and predicted single-cell cytoplasmic transcript abundance within a population of HeLa cells. See also Figure S2.

Figure 3. Cell-to-cell variability in cytoplasmic transcript abundance contains only minimal stochastic variability. (**A**) Comparison of unexplained variability ($\eta^2_{Unexplained}$) of MLR models (red), unexplained variability of three-state stochastic models (light gray), randomized data (black), Poisson limit for stochastic variability (dark blue) and the same

limit corrected for a minimal technical influence (hybridization efficiency) (light blue) in HeLa cells. (**B**) Correlation between the amount of explained variability ($\eta^2_{Explained}$) and unexplained variability ($\eta^2_{Unexplained}$) for single genes (circles) in HeLa cells. Gray area shows 90% confidence interval of a Gaussian mixture model of $\eta^2_{Explained}$ and $\eta^2_{Unexplained}$ of all genes. Blue-colored genes are outlier genes that show both the highest amount of explainable and unexplainable variability and are enriched in immediate early response genes. Red-colored genes show more explainable than unexplainable variability and are enriched in cell cycle genes. Green-colored genes are the lowest abundant genes with a spot count per cell barely above background (3-4 spots per cell), suggesting that their low level of explainable variability is mainly of a technical nature. The cross indicates a technical outlier where a discrepancy between the expression levels of the two biological replicates was observed. See also Figure S3.

Figure 4. Causality between predictors and single-cell transcript abundance. (**A**) Growing single cells on micropatterns constrains the variance of phenotypic features. Black box indicates features with a strongest reduction of variance. (**B**) Restricting the space available to single cells by micropatterns reduces cytoplasmic transcript abundance in single cells. Images show *RELA* transcripts (green) and DAPI (blue) of single constrained cells grown on differently sized micropatterns, and an unconstrained cell grown on a 10,000 μm^2 micropattern. Segmented cell outlines are shown as white lines. Scale bar is 16 μm. (**C**) Boxplots show measured single-cell spot count distributions of *RELA* transcripts in cells unconstrained on 10,000μm^2 micropatterns constrained and constrained on 300μm^2 micropatterns (n=1874 and 694, respectively). For comparison, the distribution in constrained cells as predicted by the MLR model learnt on unconstrained cells is shown, as well as a distribution arising only from Poisson noise (Pois. model). (**D**) Reduction in Kolmogorov-

Smirnov distance (*KS*) between measured single-cell transcript distributions and Poisson distributions (n = 10,000) of 9 different genes, going from unconstrained cells to constrained cells. (**E**) EGF gene-induction experiment. The heatmaps show the mean cytoplasmic transcript abundance at various time-points after serum starvation and addition of EGF as well as the *pS* of MLR models. Blue boxes highlight highest observed mean cytoplasmic transcript abundance per cell (peak expression). (**F**) The increase of *pS* from uninduced cells (red dots) to EGF-induced cells at peak expression (blue dots) follows the global trend (grey-shaded contoured area) over 583 genes that *pS* increases as transcript abundance increases in HeLa cells continuously grown in the presence of serum. (**G**) Left: Single-cell cytoplasmic transcript abundance distributions for *JUN* in uninduced cells (red) and in cells 40 minutes after EGF induction (blue). Distributions can be predicted with MLR models (black). Right: Prediction of EGF-induced cytoplasmic transcript abundance of *JUN* at the single-cell level in cell populations is achieved with MLR models learnt from a replicate experiment as well as with MLR models learnt from uninduced cells. (**H**) As (G), except for *FOS* transcripts. See also Figure S4.

Figure 5. Multi-level transcript homeostasis in single mammalian cells. (**A**) Overview of computational multiplexing. Gene-specific MLR models are applied to other genes in a pairwise manner, and genes are clustered by their similarity in being predicted by the MLR models of all other genes, resulting in a similarity matrix for both cell types. (**B**) Networks formed by connecting 2 genes (nodes) that show the 2% highest similarities. A global separation of genes based on higher correlation (log2 ratio) of cytoplasmic transcript abundance with cell area or cell volume (Figure S5D). Bar graphs show functional annotation enrichment for all genes showing a 2-fold higher correlation with cell area than with cell volume. (**C**) Genes in the networks colored according to max.-normalized correlation

between cytoplasmic transcript abundance and sets of features. Green borders indicate sub-clusters K1-3 and H1-2. (**D**) Enlargement of sub-clusters K1 and K2 present in keratinocytes. Heatmap shows max-normalized correlation between cytoplasmic transcript abundance and selected individual features. Grouping of features as in (C). See also Figure S5.

Figure 6. Nuclear compartmentalization efficiently buffers stochastic bursts in gene transcription. (**A**) Three-state stochastic modeling of gene transcript synthesis (gene module), agent-based modeling of transcript diffusion and retention in the nucleus (nucleus module), and transcript degradation (cytoplasm module). (**B**) Conceptualization of the model's output and the effect of nuclear retention on how variation between transcript synthesis events (dT_s) converts to variation between transcript export events (dT_e). (**C**) Distribution of nuclear retention times of newly synthesized transcripts of 282 genes induced by LPS in mouse bone marrow-derived macrophages. Data are derived from Bhatt et al. (Bhatt et al., 2012). Lower panels display two examples of the kinetics of the amount of chromatin-associated transcripts and transcripts in the cytoplasm during LPS induction for *FOS* and *JUN* and derived $t_{1/2}$ of gene induction, nuclear retention, and degradation. (**D**) Left, matrix of color-coded Kolmogorov-Smirnov distances (KS) of dT_s (synthesis) distributions to a Poisson distribution, as shown in (B), for multiple combinations of 'on' and 'off' times. Middle, 3 matrices of color-coded Kolmogorov-Smirnov distances (*KS*) of dT_e (export) distributions to a Poisson distribution for the combinations of 'on' and 'off' times boxed in the matrix on the left and using nuclear retention times of 15, 30 and 60 min. The white line indicating regions in the matrices where the *KS* distance is 0.1 (the region left from these lines contains KS <0.1). Right, evolution of the mean *KS* distance of dT_e distributions to a Poisson distribution over the indicated regions as a function of retention time. The dashed line indicates a specific combination of 'on' and 'off' times (5.5 min each) indicated with a

white dot in the matrix on the left relevant for immediate early genes such as *FOS* and *JUN* that show response times at the same timescale visible in (C). (**E**) The coefficient of variation (CV) of single-cell nuclear (left, black) and cytoplasmic (right, blue) transcript abundance as a function of mean transcript abundance per cell. Measured CV^2 during the EGF induction experiments are solid dots (black in the nucleus, blue in the cytoplasm), values for nuclear CV^2 obtained from the model of (A) using fitted values for cytoplasmic degradation rate and nuclear DNA content are the solid black line and grey interquantile range, and CV^2 values for cytoplasmic transcript abundance predicted with the same models are the solid blue line and blue interquantile range. (**F**) Prediction strengths (*pS*) of single-cell transcript abundance in the nucleus and cytoplasm predicted with the agent-based model in (A) (shaded area with contour lines), and with MLR models learned on measured nuclear and cytoplasmic single-cell transcript abundance (black dots). (**G**) Example time points from a 13-hour movie of doxycycline-induced HeLa 11ht MS2 cells (Halstead et al., 2015), individual transcripts are visualized using 24 MS2 stem loops and Halo-tagged MS2 coat protein. The square highlights bursts in synthesis and the arrow point to accumulation of transcripts at the nuclear envelope. (**H**) Autocorrelation function of MS2 mRNA spot counts in single cells ~1 hour after induction that show increasing cytoplasmic transcript abundance during the time of imaging. 6 optical z-planes of cells were acquired at 10 min intervals for 5 hours. N=4, data points are mean ±s.e.m. at given τ, $P<0.05$ for all $\tau>0$ min. See also Figure S6.

Figure 1

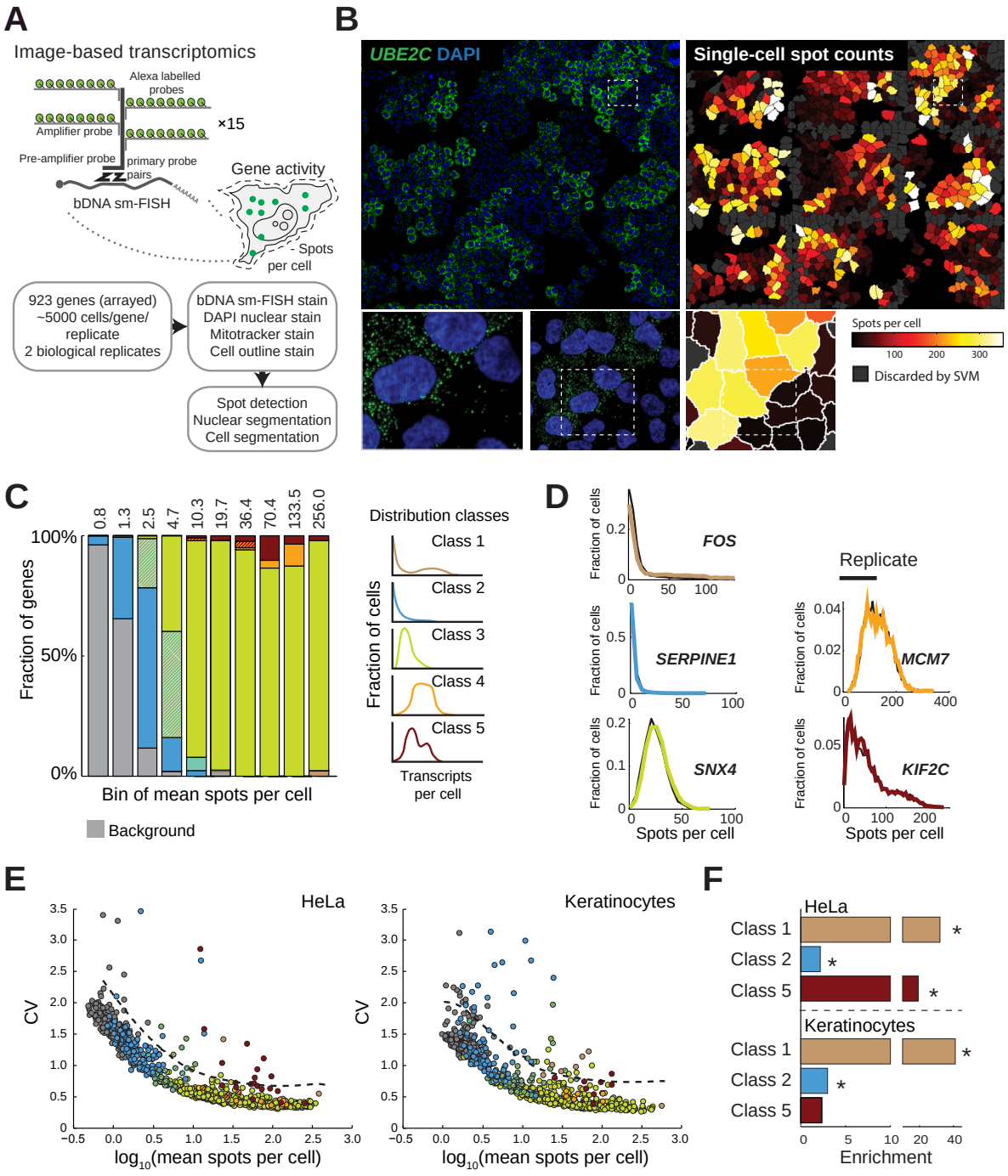

A
Image-based transcriptomics
Alexa labelled probes
×15
Amplifier probe
Pre-amplifier probe
primary probe pairs
bDNA sm-FISH
Gene activity
Spots per cell
923 genes (arrayed)
~5000 cells/gene/ replicate
2 biological replicates
bDNA sm-FISH stain
DAPI nuclear stain
Mitotracker stain
Cell outline stain
Spot detection
Nuclear segmentation
Cell segmentation
B
UBE2C DAPI
Single-cell spot counts
Spots per cell
100
200
300
Discarded by SVM
C
0.8
1.3
2.5
4.7
10.3
19.7
36.4
70.4
133.5
256.0
100%
50%
0%
Fraction of genes
Bin of mean spots per cell
Background
Distribution classes
Class 1
Class 2
Class 3
Class 4
Class 5
Fraction of cells
Transcripts per cell
D
FOS
SERPINE1
SNX4
Fraction of cells
Spots per cell
Replicate
MCM7
KIF2C
E
HeLa
Keratinocytes
CV
log10(mean spots per cell)
F
HeLa
Class 1
Class 2
Class 5
Keratinocytes
Enrichment

Figure 2

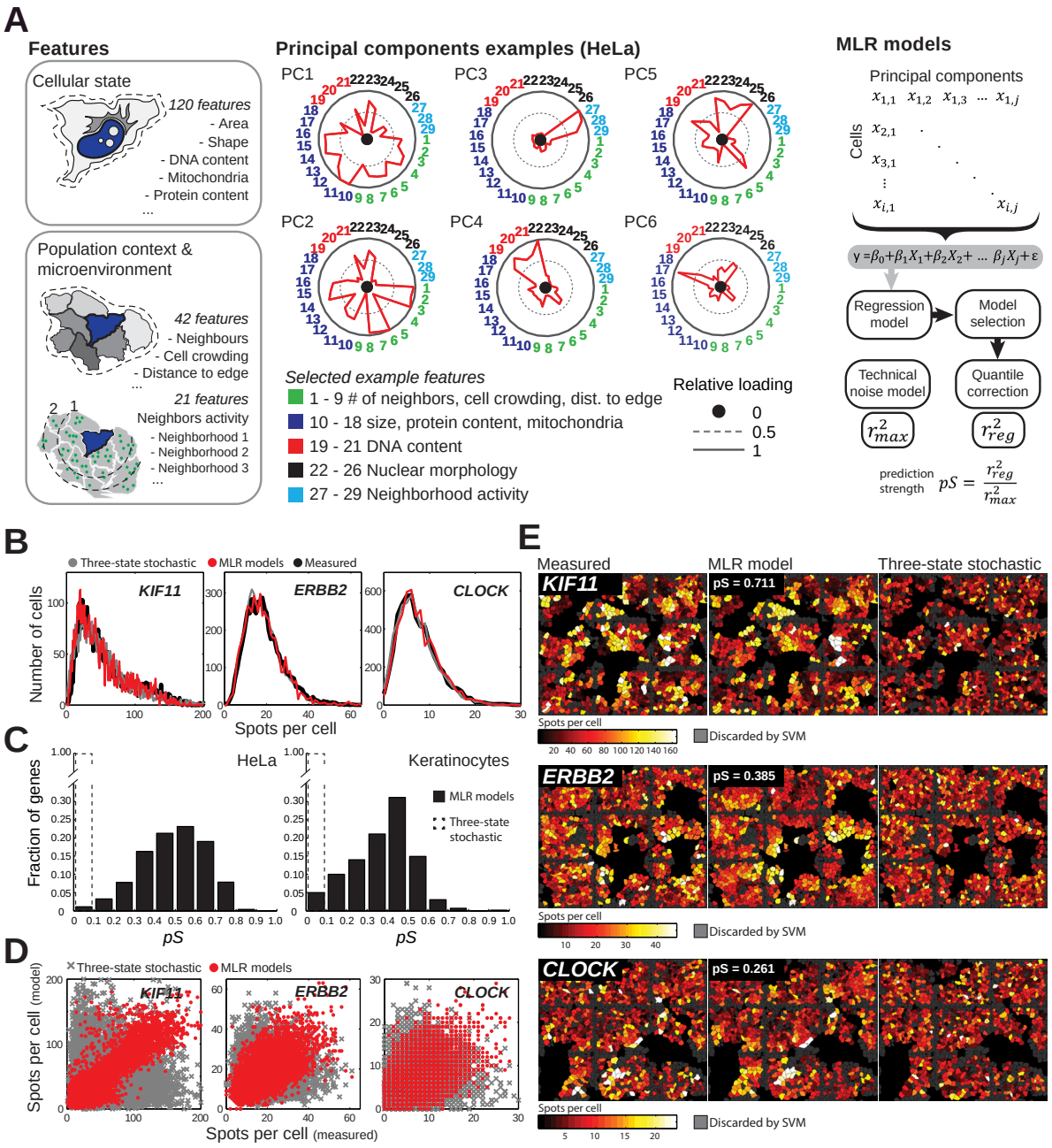


Figure 3

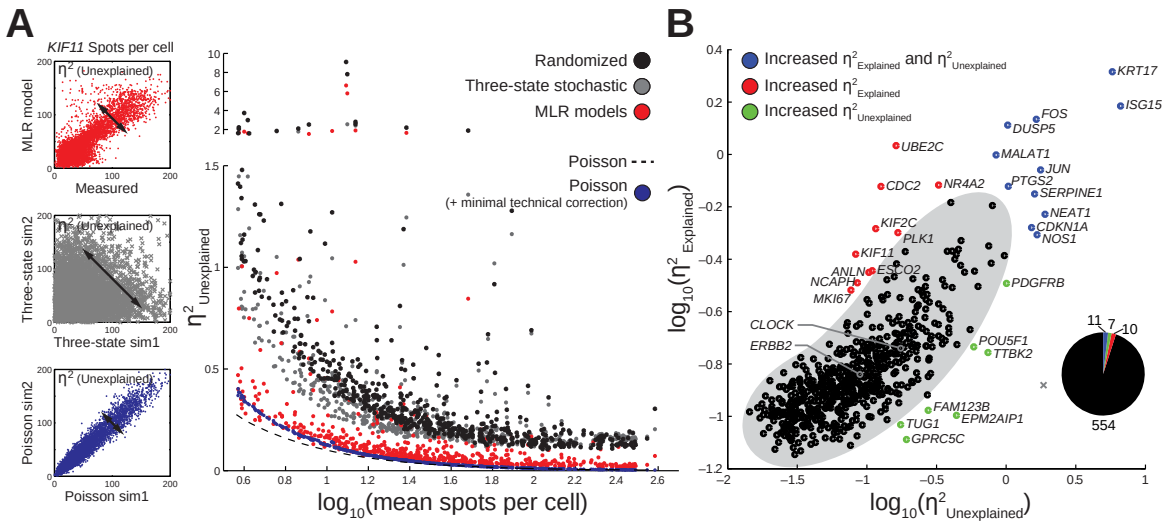


Figure 4

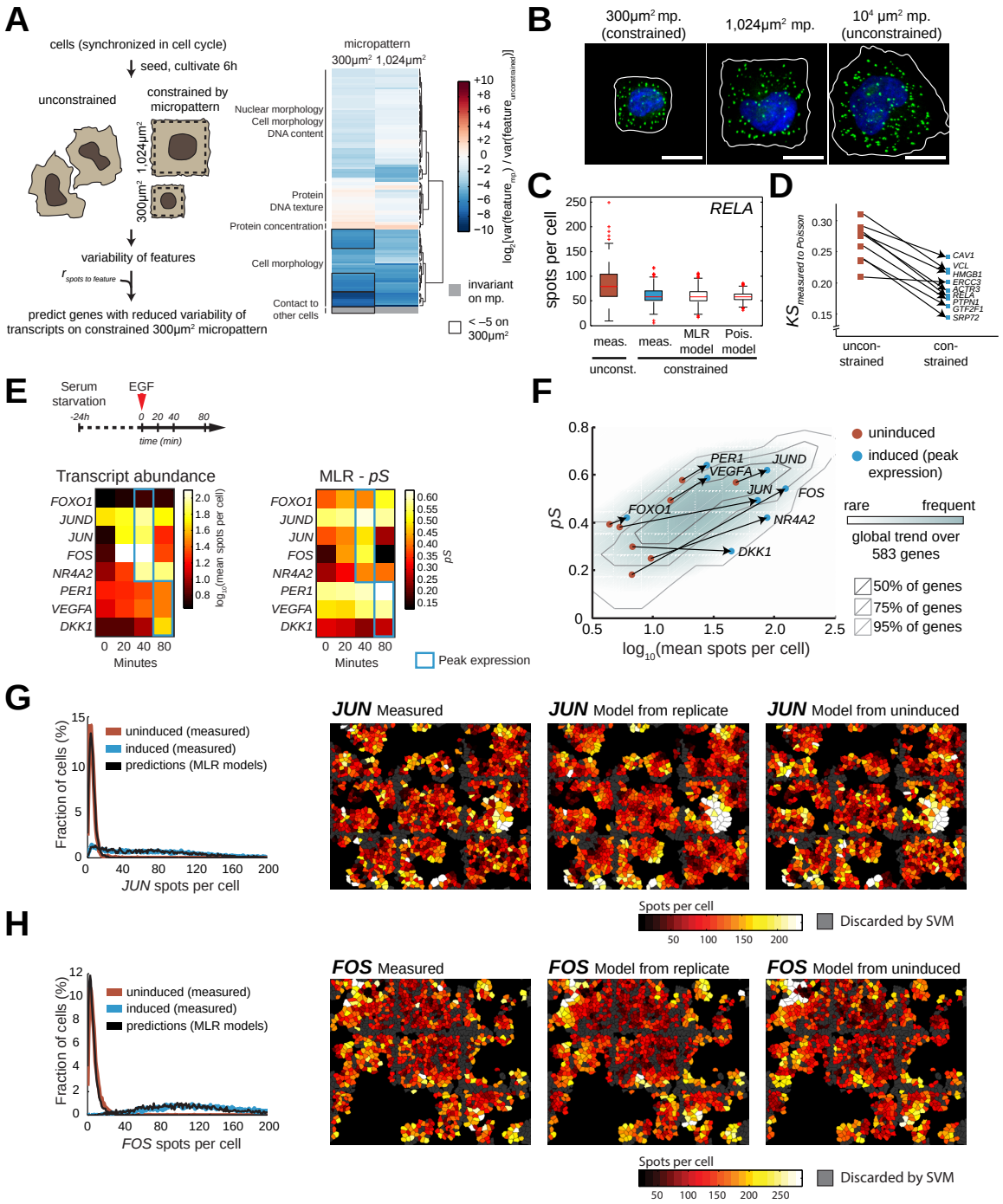

A
cells (synchronized in cell cycle)
seed, cultivate 6h
unconstrained
constrained by micropattern
variability of features
predict genes with reduced variability of transcripts on constrained 300µm² micropattern
micropattern
Nuclear morphology
Cell morphology
DNA content
Protein
DNA texture
Protein concentration
Cell morphology
Contact to other cells
invariant on mp.
< –5 on 300µm²
B
300µm² mp. (constrained)
1,024µm² mp.
10⁴ µm² mp. (unconstrained)
C
RELA
spots per cell
meas.
meas.
MLR model
Pois. model
unconst.
constrained
D
uncon-strained
con-strained
CAV1
VCL
HMGB1
ERCC3
ACTR3
RELA
PTPN1
GTF2F1
SRP72
E
Serum starvation
EGF
time (min)
Transcript abundance
MLR - pS
FOXO1
JUND
JUN
FOS
NR4A2
PER1
VEGFA
DKK1
Minutes
Peak expression
F
uninduced
induced (peak expression)
rare
frequent
global trend over 583 genes
50% of genes
75% of genes
95% of genes
G
uninduced (measured)
induced (measured)
predictions (MLR models)
JUN spots per cell
JUN Measured
JUN Model from replicate
JUN Model from uninduced
Spots per cell
Discarded by SVM
H
FOS spots per cell
FOS Measured
FOS Model from replicate
FOS Model from uninduced

Figure 5

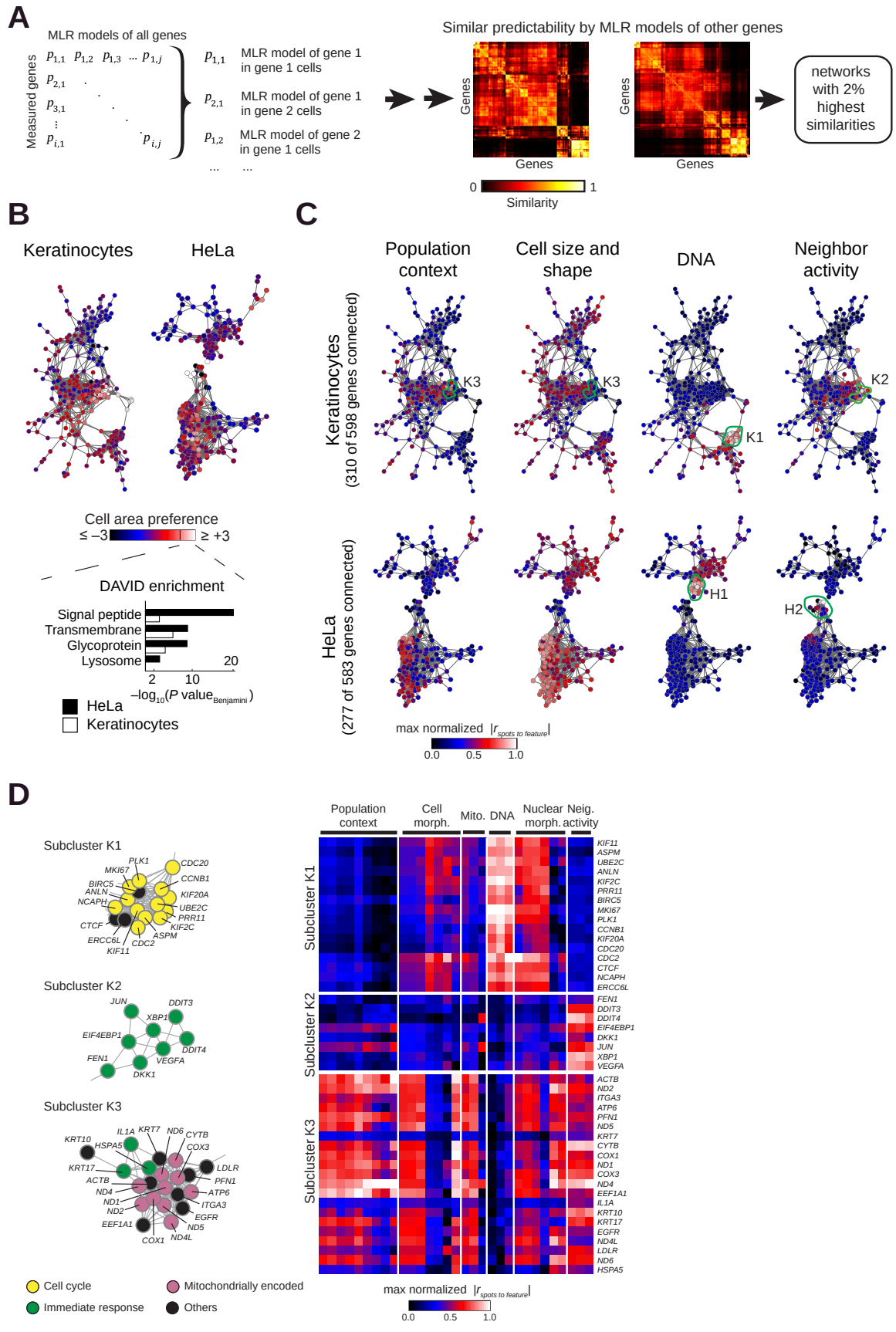
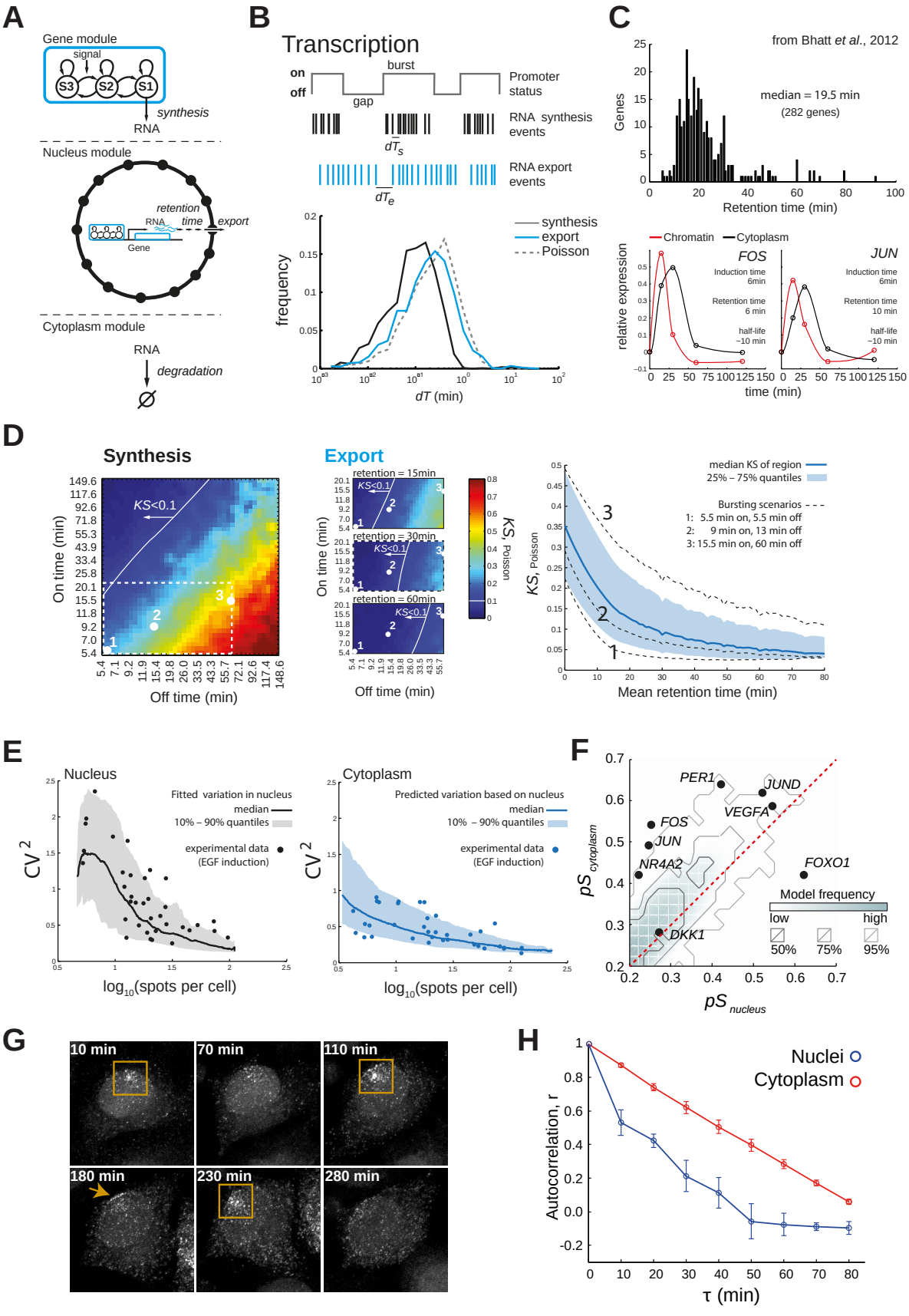


Figure 6


A
Gene module
signal
S3
S2
S1
synthesis
RNA
Nucleus module
retention
RNA
time
export
Gene
Cytoplasm module
RNA
degradation
B
Transcription
on
off
burst
gap
Promoter status
RNA synthesis events
RNA export events
synthesis
export
Poisson
frequency
dT (min)
C
from Bhatt et al., 2012
median = 19.5 min
(282 genes)
Genes
Retention time (min)
Chromatin
Cytoplasm
FOS
JUN
Induction time 6min
Retention time 6 min
Retention time 10 min
half-life ~10 min
relative expression
time (min)
D
Synthesis
Export
KS<0.1
On time (min)
Off time (min)
retention = 15min
retention = 30min
retention = 60min
KS, Poisson
median KS of region
25% – 75% quantiles
Bursting scenarios
1: 5.5 min on, 5.5 min off
2: 9 min on, 13 min off
3: 15.5 min on, 60 min off
KS, Poisson
Mean retention time (min)
E
Nucleus
Cytoplasm
Fitted variation in nucleus
Predicted variation based on nucleus
median
10% – 90% quantiles
experimental data
(EGF induction)
CV 2
log10(spots per cell)
F
PER1
JUND
VEGFA
FOS
JUN
NR4A2
FOXO1
DKK1
Model frequency
low
high
50%
75%
95%
pS cytoplasm
pS nucleus
G
10 min
70 min
110 min
180 min
230 min
280 min
H
Nuclei
Cytoplasm
Autocorrelation, r
τ (min)

Extended Computational Procedures

Classification of Distribution Shapes

Single cell distributions of false-positive wells from the same plate were subtracted from the observed distribution by randomly selecting single cells from a randomly chosen negative control cell population without gene specific probes, which was present on the same plate. For each gene and biological replicate, 5 bootstraps with replacement were performed. Distributions were classified manually using a defined set of quantitative criteria. Since single-cell distributions of endogenous human transcripts had not been characterized previously with single-molecule resolution, a human-supervised classification of every distribution was chosen to avoid a bias against unexpected single-cell distributions that might deviate from the anticipations of an automated classifier. Class 1: At least one biological replicate showed bimodality with the left mode not separating by at least 1 transcript per cell from the detection limit. Discontinuously sampled heavily skewed tails were considered to represent an additional mode. Class 2: Highest mode has 0 transcripts per cell and at least 5% of all genes have 1 transcript per cell in every bootstrap. No multimodality is observed. Class 3: Both replicates show a unimodal distribution, with the mode not being 0 transcripts per cell, and the distribution does not plateau or plateaus over less than 20% of the observed range of transcripts per cell. Class 2 to 3: At least one bootstrap of at least one replicate belongs to either class. Class 4: At least one biological replicate has a single peak, which plateaus over 20% or more of the observed range of transcripts per cell. Class 5: Transcripts show a multimodal distribution, where each peak is clearly separated from experimental background. Alternatively, the distribution has one clearly defined isolated left peak and a right shoulder, which spans at least 100% of the width of the left peak, while tolerating a sampling, which was not continuous at a bin size of one transcript per cell. Class 4 to 5: Either class 4 and class 5 are observed in distinct replicates or bootstraps or the peak declines weakly.

Feature Extraction

Area, shape, intensities and texture (at a scale of 5 pixels of individual dyes) of cells and nuclei were measured with CellProfiler (Carpenter et al., 2006), after correction of uneven illumination within single microscopic sites (Stoeger et al., 2015) and subtraction of camera dependent invariant background signal. Following analysis by CellProfiler, measurements of intensities and texture (but not other measurements like bDNA sm-FISH spots) were corrected for positional biases within each well on a per-plate basis:

coordinates of each cell within a well were binned in 30 bins along the X-coordinate and 30 bins along the Y-coordinate. Mean and standard deviation of all cells within each bin were calculated, smoothed over three bins and used to infer the position-independent measurement analogously to the z-score based correction method for single-pixel intensities described earlier. “Blue”, “Red” and “FarRed” indicate features extracted from 4,6-diamidino-2-phenylindole (DAPI), MitoTracker® Red CMXRos (Invitrogen) and Alexa Fluor® 647 carboxylic acid, succinimidyl ester (Invitrogen), respectively.

In addition to previously described measurements of the population context (Snijder et al., 2009), the number of neighbors at certain distances and the minimal distance to other cells was calculated using centroids of nuclei. Moreover, the number of directly adjacent cells and the fraction and size of extracellular space with overlap to other cells was extracted using a custom CellProfiler module, which extended the cell outline by 10 pixels. Neighbor activity was calculated by averaging the spots per cell of all cells within distinct radii surrounding the nuclei of individual cells. Note that all neighbourhood-related features were computed prior to discarding cells for data analysis (see below).

Discarding of single Cells from Data Analysis

Besides discarding every cell extending beyond the field of view (acquired image), supervised machine learning (Ramo et al., 2009) was applied to discard mitotic cells, wrongly segmented cells, multinucleate cells, and fluorescent debris (Battich et al., 2013; Stoeger et al., 2015). In addition, we discarded keratinocytes with a non-basal-like morphology.

Selection of Genes for single-cell Analysis of single Genes

The boundary of low and high expressed genes was calculated as described earlier (Battich et al., 2013), except that raw counts of spots per cell (without subtraction of false positive bDNA sm-FISH spots) were used. Genes, where the average number of bDNA sm-FISH spots per cell of both replicates was separated by three standard deviations of the boundary were used for subsequent analysis.

Automated volumetric Analysis of single Cells

Microscopic images were acquired as described above except that a step size in Z of 335nm was used. Illumination correction was performed as above using the same illumination correction statistics for all Z-layers. Volume was determined as follows using a series of custom CellProfiler modules: 1) nuclei and cells were segmented by their projection image as described above; 2) voxels with DAPI and Alexa

Fluor® 647 carboxylic acid succinimidyl ester signal above a manually chosen threshold were selected for nuclei and cells, respectively; 3) for each nucleus and cell, the largest group of 26-connected voxels above the threshold was chosen as the volume of the nucleus and cell, respectively.

Generation of the principal Components for Transcript Abundance Prediction

Features other than neighbor activity were normalized within each experimental plate by z-scoring after Winsorizing at the 0.2% and 99.8% percentiles. Features describing neighbor activity were normalized within each single well by z-scoring (to account for differences in mean gene expression between genes). Finally, each feature was normalized by z-scoring across all cells of a given cell line.

Predictions by Multi-linear regression Models

Multi-linear regression (MLR) of spot number per cell was performed using the robustfit function of MATLAB. MLR models were derived from successive additions of principal components: For each model a half of all cells measured for each gene were used as a training set, the squared Pearson correlation coefficient of the model was then calculated from a second quarter of cells for each gene. The best number of principal components was chosen when the model first reached the 0.95% percentile of the maximal squared Pearson correlation coefficient given for any number of principal components. histogram matching was done by ranking the results obtained from the linear models to the measured spots per cell from the reciprocal replicate experiments. Both results were ranked in ascending order and the assignment of the spot numbers per cell to a given cell was done by matching the relative rank of the experimental results of the replicate to the MLR results.

Predictions by Partial Least Squares Regression Models

Partial least square regression (PLSR) of spot number per cell was performed using the plsregress function of MATLAB. Models were learnt on a given data set and applied to the biological replicate. For a given well single cell data was divided into a learning set and a testing set, the testing set was used to find the number of components that achieved the maximum average r^2 from 100 bootstrap runs. These models were then applied to the biological replicate data set. As before, to obtain the final prediction of spot numbers per single cell we applied histogram matching.

Predictions by Random Forest Models

A set of 65 features describing the cellular state, population context and microenvironment (neighbor activity) were chosen manually. Models were constructed using a randomly selected subset of 50% of all cells of a single well and R's randomForest package and applied to the other half 50% of all cells of the same single well with R's predict package.

Estimation of technical Variability

To estimate the impact of technical variability on predictable variability we developed a scheme, based on conservative estimates of parameters of known technical variability of single-molecule RNA FISH. The underlying assumption of this scheme was that the observed single-cell distribution of spots per cell is a reasonable proxy of the real distribution of transcripts per cell, an assumption we have shown to be reasonable in Battich et al., 2013. Thus, we could distort this distribution with a known and defined amount of noise, representing technical variability in the cells used to build the MLR model, and in cells to which the MLR model was applied.

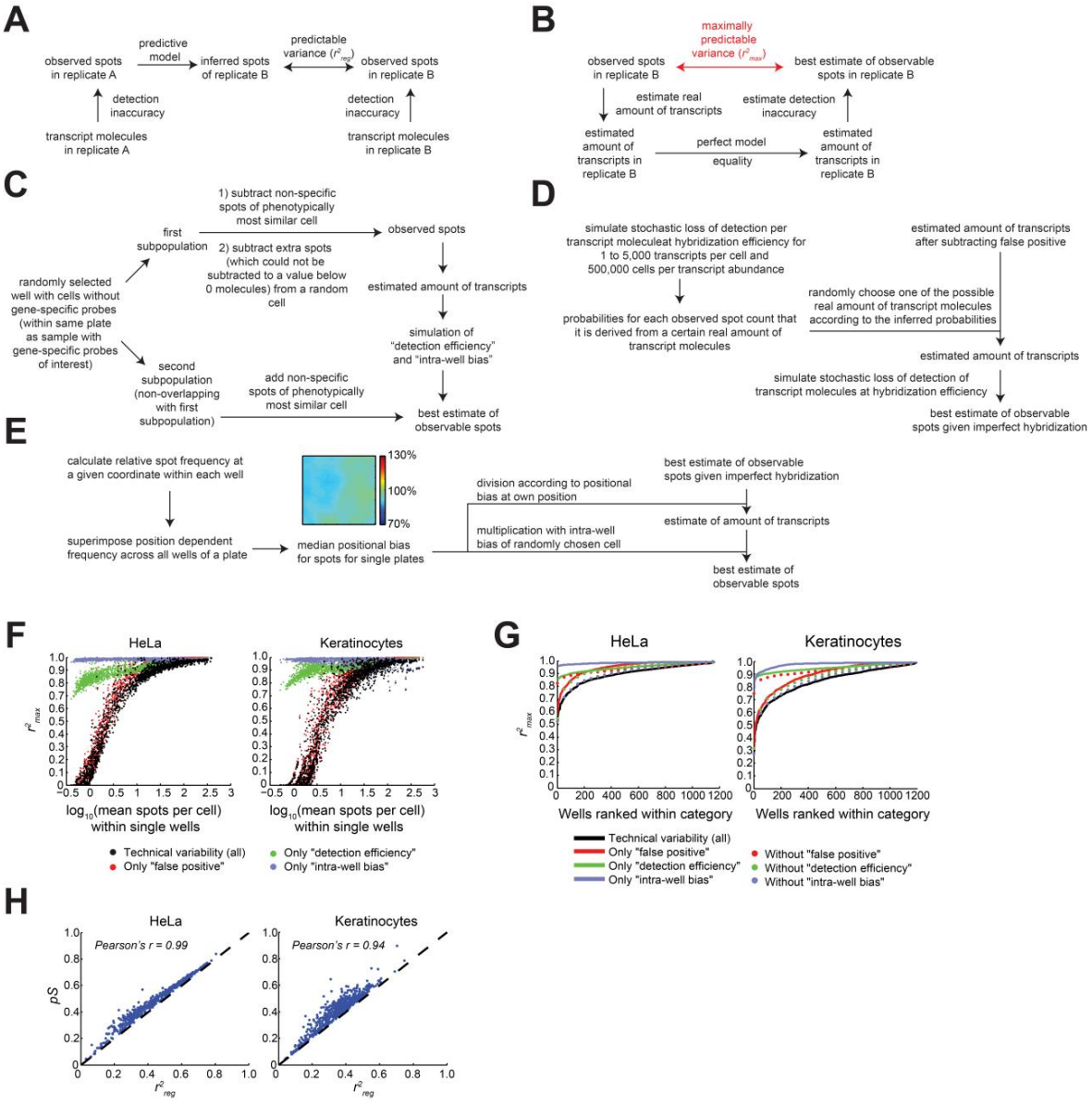
As shown in the figure below, correcting for the technical variability only has a minor effect on the predictable variability. On the other hand, we discovered that technical variability would have a stronger impact on genes with a lower mean number of transcripts per cell (see figure below). Therefore correcting for technical variability would avoid over-estimating differences between genes with a different mean number of transcripts per cell.

For each cell, 100 repetitions of the estimation of technical variability were performed with independent randomizations. Joint readouts represent the mean of Pearson's squared correlation coefficients obtained within single bootstraps.

For "false positive", a single negative control well from the same experimental plate was selected and randomly subdivided. Each cell of a given gene of interest was matched to the single cell of the negative control with the closest Euclidean distance in the phenotypic space formed by z-scored protein content and cell area. In the case that the matched cell of the negative control had a higher spot count than the cell of interest, excess spots were removed from other cells in the population of interest. At first, the excess spots of single cells were subtracted from cells which would contain at least as many remaining spots. If this was not possible, excess spots were removed randomly. Although the number of false positive spots per cell, which were independent of the presence of the targeted transcript molecules, was generally less than one or two spots per cell (Battich et al., 2013), they were the predominant source of technical variability in genes with low expression levels (see figure below).

For simulations of the “hybridization efficiency”, a detection efficiency of 85% was used, which had been determined by dual labelling experiments (Battich et al., 2013). Hybridization efficiency was simulated by assuming a maximal amount of transcript molecules per cell of 5,000 molecules (only 2 out of ~10^7 HeLa cells and 0 out of ~$4x10^6$ keratinocytes had more than 2,000 and 3,000 spots, respectively). For each possible number of transcript molecules, detection efficiency of 500,000 virtual cells was simulated by assuming detection, if a pseudorandom number from a uniform distribution between 0 and 1, which was allocated to each virtual transcript molecule, was below the indicated percentage of the hybridization efficiency. The real amount of transcripts was estimated by choosing a possible real spot count according to the probabilities that the measured transcript abundance would have been observed.

In addition to these previously known sources of technical variability in single-molecule RNA FISH, we noted a subtle positional bias in the number of transcripts at different positions within a single well (likely due to imaging artefacts). We observed this trend upon comparing the number of transcripts of single cells to the mean number of transcripts per cell of all cells within the same well, and plotting this ratio of all cells of a single plate (~200 wells) according to the spatial coordinates of every cell (see insert in figure below). While we did simulate this bias by assuming the ratio of a randomly chosen cell in the same well, we note that correcting for this “intra-well bias” had practically no effect on the predictable variability (see figure below).



Estimation of technical Error. (A) Outline of the experimental setup shows that the predictable variance (r^2_{reg}) is limited by detection inaccuracies. (B) Maximally predictable variance (r^2_{max}) is inferred by estimating the real amount of transcripts of a single cell and simulating the detection inaccuracy for the estimates of the real amount of transcript molecules. (C) Simulation of "false positive" detection events, where unspecific bDNA sm-FISH spots wrongly imply presence of additional transcript molecules. (D) Simulation of "detection efficiency" resulting from false negative detection events, where individual transcript molecules are not detected as bDNA sm-FISH spots. (E) Simulation of empirically observed "intra-well bias" at different positions within a single experimental well. (F) Maximally predictable variance considering individual known sources of technical variability. For each gene the replicate wells

of different biological replicate experiments are shown separately. (G) Maximally predictable variance of genes considered for single cell predictions. (H) Comparison of predictable variance of regression models (r^2_{reg}) and prediction strength (pS) for individual genes (dots).

Stochastic three-state Models

Stochastic RNA level simulations were carried out using the Gillespie algorithm (Gillespie, 1977) according to the following equations:

$$S3 \underset{k2}{\overset{k1}{\rightleftharpoons}} S2$$

$$S2 \underset{k4}{\overset{k3}{\rightleftharpoons}} S1$$

$$S2 \xrightarrow{k5} R$$

$$S1 \xrightarrow{k6} R$$

$$R \xrightarrow{k7} \emptyset$$

S1-S3 represent chromatin states, R is the RNA and k1-k7 are the respective reaction rate constants. Simulations were run until the coefficient of variation of the mean of the last 1000 simulation updates dropped below 0.005. Optimization of reaction rate constant parameters was done for every gene and replicate individually using the genetic algorithm from the optimization toolbox of MATLAB. The optimization function was set to minimize the relative difference in the number of cells having every given spot number in the particular experimental data and a random sample including all iterations results. The top 50 parameter sets learnt from one of the biological replicates were then applied to the replicate experiment to avoid parameter overfitting. All optimizations were carried out in the high performance computing facility of the ETH Zurich.

Distribution comparison using Kolmogorov-Smirnov Statistics

The distribution obtained by MLR models, three-state stochastic models or the reciprocal experimental replicate were compared to the measured experimental data for every gene and replicate. To calculate the Kolmogorov-Smirnov statistics, the MATLAB “kstest2” function was used.

Poisson Limit of stochastic Variability

The Poisson limit in a stochastic transcription system is found when the DNA is allowed only one state so that variation is generated only at the synthesis of RNA according to the following equations:

$$S1 \xrightarrow{k1} R$$

$$R \xrightarrow{k2} \emptyset$$

Where S1 represents the chromatin, R is the RNA and k1-k2 are the respective reaction rate constants. Simulations of this system results in a Poisson distribution with a mean given by the ratio k1/k2, which is in turn assumed to be a direct consequence of the cellular state. For every single cell in our data set we ran 10 different simulations.

In addition the initial state of the cell was assumed to be unknown and resulting from 85% hybridization efficiency, so that at the start of every minimal stochastic simulation the real spot number per cell was estimated as described for the "hybridization efficiency" in the "Estimation of technical variability". To conserve the distribution shapes and the mean expression levels observed in the experimental data, the spot number for each single cell, resulting from simulations of the lower limit of stochastic variability, was subject to an additional stochastic loss of transcripts with a probability of 0.15, thus recreating an 85% hybridization efficiency. For every single cell in our data set we ran 20 different simulations.

Quantification of unexplained Variability

Variability, which was not explained by MLR models η^2, was quantified by the procedure introduced earlier (Elowitz et al., 2002) to quantify intrinsic noise (η_{int}). Briefly, $\eta^2_{Unexplained} = \eta^2_{int} = \langle (x-y)^2 \rangle \,/\, 2\langle x \rangle \langle y \rangle$, where x and y represent two single-cell datasets being compared, and angled brackets denote mean values. For the Poisson limit, η^2 is uncorrelated variability of independent simulations at the theoretical limit of variability or adding simulated noise given 85% hybridization efficiency as discussed above. This was computed for each gene, taking the mean of intrinsic noise calculations for all possible pairwise combinations of the simulation results (as described above). For the unexplained variability of MLR models, η^2 was calculated using the experimentally measured spot numbers and the prediction from the model learned from the reciprocal replicate. For the Three-state stochastic model models, unexplained variability was calculated using two different simulation runs of the

models. Unexplained variability after randomization, was calculated with randomized results from the stochastic limit simulations with added technical noise.

Quantification of explained Variability

Variability, which was explained by MLR models ($\eta^2_{Explained}$), was quantified by the procedure introduced earlier (Elowitz et al., 2002) to quantify extrinsic noise (η_{ext}). Briefly, $\eta^2_{Explained} = \eta^2_{exp} = (\langle xy\rangle - \langle x\rangle\langle y\rangle)/(\langle x\rangle\langle y\rangle)$, where x and y represent two single-cell datasets being compared, and angled brackets denote mean values (See Supplementary Software).

Bayesian Network Inference

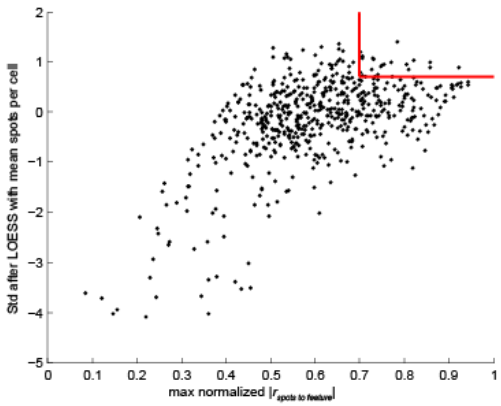
Bayesian network inference was performed with the Bayesian Network Toolbox for MATLAB (Kevin Murphy, http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html) and the BNT Structure Learning Package for MATLAB (Philippe Leray and Olivier Francois, http://bnt.insa-rouen.fr/programmes/BNT_StructureLearning_v1.3.pdf). For the nascent transcript dataset single cell measurements of nascent transcripts in the cytoplasm of cells, local cell density, cell area and protein and DNA content were combined with measurements of the population size (number of cells in a well), and the number of cells seeded. For the HeLa cells image-based transcriptomics dataset Bayesian network inference was performed in dependently for the two biological replicates using the single cell measurements of transcript abundance (spots per cell), local cell density, cell area and protein and DNA content. All measurements were discretized maximizing the Akaike Information Criterion, as implemented in the Structure Learning Package and Bayesian network inference performed using a Monte Carlo Markov Chain search over directed acyclic graphs (DAGs) assuming fully observed data. The procedure was boostrapped 2,000 times by sampling with replacement 33,000 cells in the nascent transcript dataset and 66% of all cell in the image based transcriptomics dataset. Edges where the average weight was less than 0.05 (i.e. they appeared in less than 5% of all Bayesian network inference runs) were discarded. In addition, for edges with unresolved directionality the lowest raking directionality was discarded only if the difference to the highest ranked directionality was of 20% or higher.

For the single gene analysis, only reproducible directed edges were shown. In addition, nodes were considered as having no incoming edges only when in both replicates there were only outgoing, undirected or inferred edges connecting it to all other nodes.

Reduction of Variability on Micropatterns

For analysis of cells grown on 300 μm^2 and 1,024 μm^2 micropatterns, cells were only included if they adapted their shape to the micropattern, and if they did not touch another cell. Single-cell features were measured as described above, except that a mitochondrial stain had been omitted. Features describing the neighbor activity, and most features describing the population context of single cells, were excluded from analysis since the spacing between cells was given by the layout of the micropatterns on the cover slip, and were thus invariant.

Prediction of genes with reduced transcript variability in cells grown on 300 μm^2 micropatterns was based on 32 features whose variance was reduced to at least 10^{-5} of the variance seen amongst unconstrained cells that were grown on 10,000 μm^2 micropatterns (and on which at least 25% of the area was not occupied by cells). For all genes the max.-normalized correlation between spots and features and their LOESS regressions against the mean number of transcripts per cell were used. The values of the 32 different features were averaged, and genes, with a mean max.-normalized correlation above 0.7 and a mean to the LOESS fit by 0.7 standard deviations were considered for analysis (see figure below). Of the 22 possible candidates, 9 genes were chosen manually to reflect different biological functions.



Prediction of Spots per Cell of EGF inducible Genes

A set of 45 features, which were chosen to represent different properties of the cellular state, population context and microenvironment, were Winsorized at the 0.2% percentiles and normalized by z-scoring across the full plate, except for neighbour activity features, which were normalized by z-scoring within the single well. All normalized features were used directly for MLR models without selection of principal components.

For predictions of "uninduced" and "induced" spots per cell, MLR models of biological replicate experiments were applied and predictions histogram-matched to the single-cell spot distribution of the cell population used to construct the MLR (Model from replicate). For predictions of induced transcripts by the MLR model of an uninduced cell population, the predictions were histogram-matched to the observed single-cell spot distribution of an induced biological replicate (Model from uninduced). For predictions of the quasi steady-state, MLR models were applied to a replicate well, which was present on the same multi-well plate and predicted spots per cell histogram-matched to the single-cell spot distribution of the cell population used to construct the MLR.

Clustering by mutual Prediction

For every gene, the MLR model was applied to all cell populations of the biological replicate experiment. Thus for every population of cells, in which the transcripts of a given gene had been measured, there would be predictions by the gene-specific MLR models of all other genes. For each gene and biological replicate, the prediction strengths by the multi-linear regression models of all genes of the second biological replicate were determined without histogram matching. For each gene whose transcripts were predicted, the prediction strengths resulting from all models were normalized by division with the mean prediction strength for this gene. For each biological replicate, the connection specificity index was calculated as described before using the default parameter value (Green et al., 2011). For the network visualization individual gene-to-gene edges are shown if the average connection specificity index of both replicates was within the 2% strongest connection specificity indices. Networks were visualized in Cytoscape (Shannon et al., 2003).

Functional Enrichment Analysis

Functional enrichment analysis was performed with DAVID v6.7(Huang et al., 2009).

Signature Heatmaps

The signature heatmaps for the gene networks were derived from the absolute correlation values of a given feature to the spot count of the cell, normalized by the maximum absolute correlation value of that feature and all genes.

Estimations of Transcript Retention Times from Bhatt *et al.*

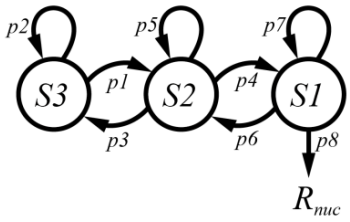
The response dynamics of transcripts associated with the chromatin compared to transcripts in the cytoplasm of mice macrophages in response to LPS were obtained from Bhatt *et al.*, 2012 (Bhatt et al., 2012) (GEO accession identifier GSE32916, data corresponding to figure 3 in Bhatt et al 2012). Response curves of every gene included multiple samples obtained at different time points after LPS stimulation. For every gene, transcripts associated with the chromatin, and transcripts in the cytoplasm, response curves were independently normalized by the sum of FPKM values in the curve. Only response curves that showed an increase in expression in both chromatin and cytoplasm were considered for further analysis. From data points response curves were interpolated using the cubic interpolation tool of the *interp1.m* function of MATLAB, to account for RNA synthesis time, and to avoid underestimation of nuclear retention time. All genes, whose interpolated response curves of chromatin-associated and cytoplasmic transcripts crossed before the peak of induction, were discarded. The interpolated line in the cytoplasm was not allowed to increase for the first minute in the response curve. Then, we defined $\Delta t_{th} = t_{cyt,th} - t_{chr,th}$ where $t_{cyt,th}$ is the estimated time at which the interpolated cytoplasmic response curve first exceeds the normalized expression threshold (th), similarly $t_{chr,th}$ is the estimated time at which the interpolated chromatin response curve first exceeds th. We varied th between 0.1 and 0.2, and defined the retention time in the nucleus as the maximal Δt_{th} observed.

Agent-based modeling of transcript synthesis and export

To test whether a nuclear compartment with physiologically relevant properties would buffer cytoplasmic transcript abundance we built a multilayer agent base model that recapitulate key parameters in the life of an mRNA, and we run simulations using a time resolution of 100ms and experimentally measured parameters from the literature whenever possible. The model consists of three main modules, or parts. Briefly, the Gene Module is a three-state stochastic process responsible for the synthesis of mRNA. The Nuclear Module simulates the 3D nuclear compartment that contains the Gene Module at a given location, it allows free diffusion of mRNA, and also contains NPs at its boundaries and allows the interaction of mRNA with the nuclear pore complexes (NPCs). Finally, mRNA is degraded in the Cytoplasm Module with a single step probabilistic decay rate. Source code for the model can be found in the Supplementary Software.

The Gene Module

The Gene Module simulates the following probabilistic process:



Where $S1\ to\ S3$ are mutually exclusive states of the gene, p_{1-8} are the transition probabilities between the states $S1\ to\ S3$ and sink event to produce an mRNA molecule in the nucleus R_{nuc}. Decisions of transition takes place on every time lapse update of the model (time lapse of simulation, $\tau = 100ms$ for all simulations), such that the sum of corresponding transition probabilities from a given state always sums up to 1.

To allow simulation of induction experiments and the dependence of RNA synthesis with total protein content $p_1$and $p_4$were set as functions of the induction time and cellular protein content respectively. Thus,

$$p_1 = \frac{k_1}{b^{\Delta t * a}}\ , and\ p_1 + p_2 = 1$$

$$p_4 = k_2 * \left(1 - \frac{1}{b^{PC * a}}\right)\ , and\ p_3 +\ p_4 + p_5 = 1$$

where, $\Delta t = |t_i - t_u|$, t_i is the current simulation time, t_u is the gene induction time, k_1 is the raw transition probability between $S3$ and $S2$, k_2 is the raw transition probability between $S2$ and $S1$, PC is the protein content of the cell (measured integrated intensity of succinimidyl ester staining divided by 100), and a and b are scaling factors (see Tables S2-3, for tested range on these parameters).

The Nuclear Module

The nucleus was modelled as a sphere or radius r_n that contained a number of genes $N_g$at location $G_{x,y,z}$. During simulations N_g could vary between 1 and 5 depending on the measured DNA content of the cells (0.36% $N_g = 1$, 59.56% $N_g = 2$, 16.00% $N_g = 3$, 23.23% $N_g = 4$, and 0.86% $N_g = 5$), note that the majority of the cells have between 2 and 4 gene copies, and gene copies of 1 or 5 occurs only in ~1.2% outlier cells. At the surface of the sphere the nucleus has a number of NPC, N_{npc}, located randomly. Newly synthesised R_{nuc} agents can diffuse freely with in the sphere following the formula, $dx, dy, or\ dz = nrand(\ \ ) * \sqrt{dD\tau}$, where d is the number of dimensions (3), D is the diffusion

coefficient, here set to 0.004 $\mu m^2 s^{-1}$ for all simulations (Mor et al., 2010), τ is the time lapse update, and $nrand(\ \ )$ is the output of *randn()* MATLAB function.

Interaction with the NPCs was assumed to be possible only if an mRNA R_{nuc} was 60nm from the NPC center, approximately the radius of the NPCs (Stuwe et al., 2015). Upon interaction of an R_{nuc} object with an NPC the probability of transport to the cytoplasm p_9 was of 0.3 for all simulations, which is close to measured transport probabilities (Grunwald and Singer, 2010). If, an R_{nuc} would fail to be transported its position was returned to the position prior interaction with the NPC.

The Cytoplasmic Module

The cytoplasmic module catalyzed the simple reaction:

$$R_{cyt} \xrightarrow{p_{10}} \emptyset$$

Where R_{cyt} is the RNA in the cytoplasm, and p_{10} is the probability for R_{cyt} to be degraded within a period of time τ.

Simulations for quantification of buffering strength

To reduce computation time, the Gene Module and the Nuclear Module were first simulated separately with parameters range as shown in Table S1, resulting in physiological relevant times for burst, gaps and nuclear retention time of mRNA. For every synthesized R_{nuc} its retention time was sampled from recomputed retention times using the Nuclear Module (Note, at this stage we do not yet consider the degradation rate). Simulations were run using $\tau = 100ms$ for a total of 24-hours, p_1 was set to 1 and $p_4 = k_2$.

To quantify the buffering effect on the synthesis of RNA directly we defined dT_s as the set of absolute differences between the timing of a synthesis event $s_n$and s_{n-1}. Similarly, dT_e was defined as the set of absolute differences between the timing of a export event $e_n$and e_{n-1},where $e = s + Ret$, s being the synthesis event and the Ret the sampled retention time. The respective Poisson process for the given simulation run was defined as having the same amount of synthesis or export events within the 24-hours simulation period, but the events were randomly distributed over time (dT_p). Then the distribution of dT_s or dT_e were compared to the respective dT_p distribution using Kolmogorov-Smirnov statistics.

Parameters for quantification of buffering strength	
Parameter	Value range
p_1	1
p_4	10^{-5} – $10^{-3.5}$
p_6	10^{-5} – $10^{-3.5}$
p_8	$5x10^{-3}$
p_{10}	NA
N_{npc}	500 - 10^4
N_g	1
r_n	5µm - 10µm
Note: 100 cells were simulated per parameter combination. 1600 different parameter combinations	

Simulations of induction experiments and prediction of cytoplasmic variability

Retention times of R_{nuc} were pre-computed resulting in a mean retention time of ~18 min. Then simulations of RNA synthesis, export and degradations were done with random sampled parameters within ranges given below. Simulations where run for 80 min using $\tau = 100ms$ and $p_4 = k_2$. The final simulations were sampled for R_{nuc} and R_{cyt} readouts at 20, 40 and 80 min.

To predict variability in the cytoplasm we selected models that at given time point showed a close relationship between the mean R_{nuc} counts and the variability of R_{nuc} counts compared to that measured in the EGF induction experiment for nuclear transcripts. Then we computed the corresponding values that the simulation gave using R_{cyt} mean counts and variability and compared to cytoplasmic transcript abundance and variability in the EGF induction experiment.

Parameters for predictability change between nucleus and cytoplasm	
Parameter	Value range
k_1	10^{-7} – $5x10^{-3}$
p_2	10^{-7} – $5x10^{-3}$
p_4	10^{-7} – $5x10^{-3}$
p_6	$2x10^{-3}$ – $8x10^{-3}$
p_8	10^{-4} $5x10^{-1}$
$a\ (p_1)$	0 – $5x10^{-3}$
$b\ (p_1)$	1.1 – 2
t_i	0 – 80min
p_{10}	10^{-7} – $5x10^{-3}$
N_{npc}	500 - 10^4
N_g	1 – 5
r_n	5µm - 10µm
Note: 500 cells were simulated per parameter combination. ~30000 different parameter combinations	

Simulations for quantification of predictability change between nucleus and cytoplasm

Simulations were run as before but allowing p_4 to scale with the protein content of single cells observed in a population of 5593 HeLa cells. To quantify predictability at peak induction of R_{nuc} vs R_{cyt} we learned MLR models on the simulated data as descried before for the experimental data. In this case the prediction strength is equivalent to the r^2 value.

Parameters for predictability change between nucleus and cytoplasm	
Parameter	Value range
k_1	10^{-7} – $5x10^{-3}$
p_2	10^{-7} – $5x10^{-3}$
k_2	10^{-7} – $5x10^{-3}$

p_6	$2x10^{-3} – 8x10^{-3}$
p_8	10^{-4} $5x10^{-1}$
$a\ (p_1)$	$0 – 5x10^{-3}$
$b\ (p_1)$	1.5 – 2
$a\ (p_4)$	0.1 – 10
$b\ (p_4)$	1.01
t_i	0 – 50min
p_{10}	$5x10^{-9} – 5x10^{-3}$
N_{npc}	500 - 10^4
N_g	1 – 5
r_n	5µm - 10µm
Note: 400 cells were simulated per parameter combination. ~25000 different parameter combinations	

References

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., *et al.* (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome biology *7*, R100.

Gillespie, D.T. (1977). Exact Stochastic Simulation of Coupled Chemical-Reactions. J Phys Chem-Us *81*, 2340-2361.

Green, R.A., Kao, H.L., Audhya, A., Arur, S., Mayers, J.R., Fridolfsson, H.N., Schulman, M., Schloissnig, S., Niessen, S., Laband, K., *et al.* (2011). A high-resolution C. elegans essential gene network based on phenotypic profiling of a complex tissue. Cell *145*, 470-482.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.

Mor, A., Suliman, S., Ben-Yishay, R., Yunger, S., Brody, Y., and Shav-Tal, Y. (2010). Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells. Nat Cell Biol *12*, 543-552.

Ramo, P., Sacher, R., Snijder, B., Begemann, B., and Pelkmans, L. (2009). CellClassifier: supervised learning of cellular phenotypes. Bioinformatics *25*, 3028-3030.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.

Stuwe, T., Correia, A.R., Lin, D.H., Paduch, M., Lu, V.T., Kossiakoff, A.A., and Hoelz, A. (2015). Nuclear pores. Architecture of the nuclear pore complex coat. Science *347*, 1148-1152.

Extended Experimental Procedures

Measurement of Nascent Transcripts

Nascent transcripts were detected with Click-iT RNA 488 Imaging Kit (Invitrogen) after 1h of incubation with 5-ethynyl uridine. The original protocol of the manufacturer was adapted for processing with residual volume (to prevent cellular detachment) by leaving residual volumes of 15µl PBS and adding 15µl of 2× reagents. Total cytoplasmic fluorescence was quantified by integrating the intensities of cytoplasmic pixels after illumination correction and background subtraction.

Micropatterns

For measuring variability of features of cells on micropatterns, HeLa cells were grown at 80% confluence, cell-cycle synchronized by mitotic shake-off and propagated for two days. Then cells were again isolated by mitotic shake-off and seeded to a culture dish. After 3.5h single cells were isolated by trypsinization and 30,000 cells seeded on a PADO1-SQRS cover slip (Cytoo) and fixed after 6h. bDNA sm-FISH staining procedures was performed as in image-based transcriptomics, except that no protease was added to avoid detachment of cells. MYC and HPRT1 transcripts, were detected in parallel by Type1-488 and Type6-650 ViewRNA Signal Amplification Kits (Affymetrix). Only cells, which had fully spread on the 300 μm^2 and 1,024 μm^2 patterns were used for analysis. Cells seeded onto 10,000 μm^2 patterns were not constrained by the size of the pattern. Images were acquired on a Nikon Eclipse Ti Spinning Disk microscope using a CSU-W1-T2 disk and a 0.95NA 40x objective (Nikon). For measuring the variability of multiple different transcripts on micropatterns, we adopted the upper protocol by substituting the mitotic shake-off with computational gating for singular DNA content, and imaging in an automated CellVoyager 7000 (Yokogawa) using a custom adapter (Yokogawa) for Superfrost Excell slides (Menzel-Gläser).

RNA Interference Screens

We used measurements for nuclear area and local cell density from druggable genome screens that had been performed previously in our lab to assay the infection of HeLa cells by SV40, MHV, VV, VSV, HPV16 and HSV1 (Snijder et al., 2012). Phenotypic strengths are in the absence of any population context correction (Snijder et al., 2012).

EGF Induction

HeLa cells were grown in DMEM+FBS for 3 days, washed twice with PBS and three times with and cultivated in DMEM without serum for 24h (serum starvation). EGF (Millipore) was added in serum free medium to establish a final concentration of 20ng/ml. For measurements of the quasi-steady state (in the presence of serum), cells were fixed and processed after 3 days of growth in DMEM+FBS. To preclude a potential effect of MitoTracker® Red CMXRos on immediate response genes, no MitoTracker was added to "unstimulated", "stimulated" and "quasi-steady state" measurements of EGF inducible transcripts.

Detection of Nuclear Transcripts

Glacial acetic acid was added during the fixation at a final concentration of 2.5% [v/v]. The number of transcripts at the transcription site was determined by dividing the total intensity of the burst by the average intensity of single spots.

Supplemental Figure Legends

Figure S1. Image-based transcriptomics of cell-to-cell variability in cytoplasmic transcript abundance, Related to Figure 1. (**A**) A keratinocyte cell population stained for cytoplasmic *UBE2C* transcripts (bDNA sm-FISH in green). Visualization of the quantified cytoplasmic transcript abundance (spots per cell) by pseudo-coloring single-cell segmentations. Dashed boxes mark enlargements. Cells are discarded by machine learning (SVM, grey) when they touch image borders or are wrongly segmented or show signs of differentiation. (**B**) Classification of single-cell distributions of cytoplasmic transcript abundance in keratinocytes cells. The mean spots per cell for the genes in each bin is indicated on top.

Figure S2. Predicting cytoplasmic transcript abundance in single cells within a population, Related to Figure 2. (**A**) Selection of genes on the basis of their mean spots per cell (see also Battich et al., 2013). (**B**) Correlation between transcript abundance and single-cell features. (**C**) Left side: Cells binned by protein content and cell volume. Light gray shows 25- and 75-percentiles, black line median. Right side: single-cell correlation between protein content and features describing nuclear and cellular volume and area. (**D**) Relative loading of the first 6 principal components for 29 representative features. (**E**) Kolmogorov-Smirnov statistic comparing the measured shape of single-cell transcript distributions to the biological replicate experiment and the predictions by multi-linear regression (MLR) models and three-state-stochastic models (3SS). (**F**) Prediction of single-cell transcript distributions of *KIF11*, *ERBB2* and *CLOCK* in keratinocytes by MLR models and three-state stochastic models. (**G**) Upper part: Prediction of *KIF11*, *ERBB2*, and *CLOCK* cytoplasmic transcript abundance in single keratinocytes by MLR models and three-state stochastic models. Lower part: Visualization of measured and predicted single-cell cytoplasmic transcript abundance within a population of keratinocytes. (**H**) Global trend between mean spots per cell and prediction strengths (*pS*).

Figure S3. Cell-to-cell variability in cytoplasmic transcript abundance contains only minimal stochastic variability, Related to Figure 3. (**A**) Comparison of unexplained variability ($\eta^2_{Unexplained}$) of Multi-linear regression (MLR), Partial Least Squares (PLSR), and Random Forest models (red), unexplained variability of three-state stochastic models (light gray), randomized data (dark gray) and Poissonian variability with minimal technical error (blue) and without technical error (dashed line) for single genes (circles) in HeLa cells. (**B**) Comparison of unexplained variability ($\eta^2_{Unexplained}$) of MLR models (red), unexplained variability of three-state stochastic models (light gray), randomized data (dark gray) and Poissonian variability with minimal technical error (blue) and without technical error (dashed line) for single genes (circles) in keratinocytes. (**C**) Correlation between the amount of explained

variability ($\eta^2_{Explained}$) and unexplained variability ($\eta^2_{Unexplained}$) for single genes (circles) in HeLa cells. Similar to Figure 3B, except that genes are here additionally color-coded for mean cytoplasmic transcript abundance in single cells. Cross indicates technical outlier.

Figure S4. Causality between predictors and single-cell transcript abundance, Related to Figure 4. (**A**) Outline of Bayesian network inference analysis between predictors and gene-specific transcripts in HeLa cells. (**B**) Bayesian network inference places transcript abundance downstream of predictors for 83% of genes and downstream (in between) for 17% of genes in HeLa cells. (**C**) Closer analysis of the 77 in between genes in (B), including examples of genes whose cytoplasmic transcript abundance lies downstream of cell area and protein content and upstream of DNA content (*CDK1* and *POLA1*), or upstream of cell crowding (*ACTR2* and *3*, encoding the Arp2/3 complex, and *RHOA*). (**D**) Upper part: Bayesian network inference on bulk transcription rate, measured by 1h of 5-ethynyl uridine incorporation, after experimentally varying the amount of cells seeded into a single well (gray circle). Lower part: Occurrence of directed edges among all Bayesian networks inferred in (A). (**E**) Plots of the median z-scored RNAi effect (n=54: 3 siRNAs per gene, 6 biological, and 3 technical replicates) for 367 genes on two single-cell features (cell crowding and nuclear area) against the correlations that these features show with cytoplasmic transcript abundance of the same genes. Red dots represent genes where silencing led to a reduction of the mean number of cells by 25% or more. (**F**) *pS* increases from cells continuously grown in the presence of serum (magenta dots, quasi-steady state) to EGF-induced cells at peak expression (blue dots) and falls into the global trend (grey-shaded contoured area) over 583 genes that *pS* increases as transcript abundance increases in HeLa cells continuously grown in the presence of serum. (**G**) Prediction of *JUN* and *FOS* cytoplasmic transcript abundance in single HeLa cells by MLR models before addition of EGF (uninduced) and 40 minutes after adding EGF (induced).

Figure S5. Multi-level transcript homeostasis in single cells, Related to Figure 5. (**A**) Construction of gene similarity network from similarity matrix, exemplified for keratinocytes. (**B**) Enlargement of similarity matrix with genes shown in network of keratinocytes. K1, K2, K3 indicate clusters corresponding to sub-clusters of network. (**C**) As in (B), except for HeLa cells. (**D**) A global separation of genes based on higher correlation (log2 ratio) of cytoplasmic transcript abundance with cell area or cell volume. Cell area preference is defined as the ratio of the correlation of spots to cell area and to volume - given in standard deviations after LOESS fit to mean spots per cell. (**E**) Enlargement of sub-clusters H1 and H2 present in HeLa cells. Heatmap shows max-normalized correlation between cytoplasmic transcript abundance and selected individual features. Grouping of features as in Figure 5C. Note that

cytoplasmic abundance of the predominantly nuclear MALAT1 and NEAT1 transcripts is increased ~5 fold after the open mitosis of human cell divisions (not shown).

Figure S6, Related to Figure 6. Nuclear compartmentalization efficiently buffers stochastic bursts in gene transcription. (**A**) 3 matrices of color-coded Kolmogorov-Smirnov distances (*KS*) of dT_e (export) distributions to a Poisson distribution for the full range of combinations of 'on' and 'off' times as in the left panel of Figure 6D, using nuclear retention times of 15, 30 and 60 min. The white line indicates regions in the matrices where the *KS* distance is 0.1 (the region left from these lines contains *KS* <0.1). (**B**) The heatmaps show the mean transcript abundance at various time-points after serum starvation and addition of EGF for nuclear transcripts (left side) and cytoplasmic transcripts (right sight). Blue boxes highlight highest observed mean nuclear and cytoplasmic transcript abundance per cell (peak expression). (**C**) Panels show cells stained for nuclear and cytoplasmic *NR4A2* transcripts (green) at indicated time-points after EGF induction. Nuclear *NR4A2* transcripts were made accessible to bDNA sm-FISH by including acetic acid during the fixation of the cells. Nuclei were stained with DAPI (magenta). (**D**) Expression over time and autocorrelation function of TRICK mRNA spot counts in individual HeLa 11ht IRT cells ~1 hour after induction. Movies were recorded for 5 hours at 10 min resolution, sampling 6 z-planes per time-point.

Supplemental Movie Legend

Supplemental Movie 1. TRICK mRNA in individual HeLa, Related to Figure 6. Timestamp shows minutes passed since the start of the image acquisition.

Supplemental Table Legend

Supplemental Table 1. Similarity of genes in computational multiplexing, Related to Figure 5.

Supplemental File Legend

Supplemental File 1. Cytoscape networks of computational multiplexing, Related to Figure 5. Networks formed by connecting 2 genes (nodes) that show the 2% highest similarities. A global separation of genes based on higher correlation (log2 ratio) of cytoplasmic transcript abundance with cell area or cell volume.

Supplemental Software Legend

Supplemental Software 1. Agent based model of transcription, Related to Figure 6.

Figure S1

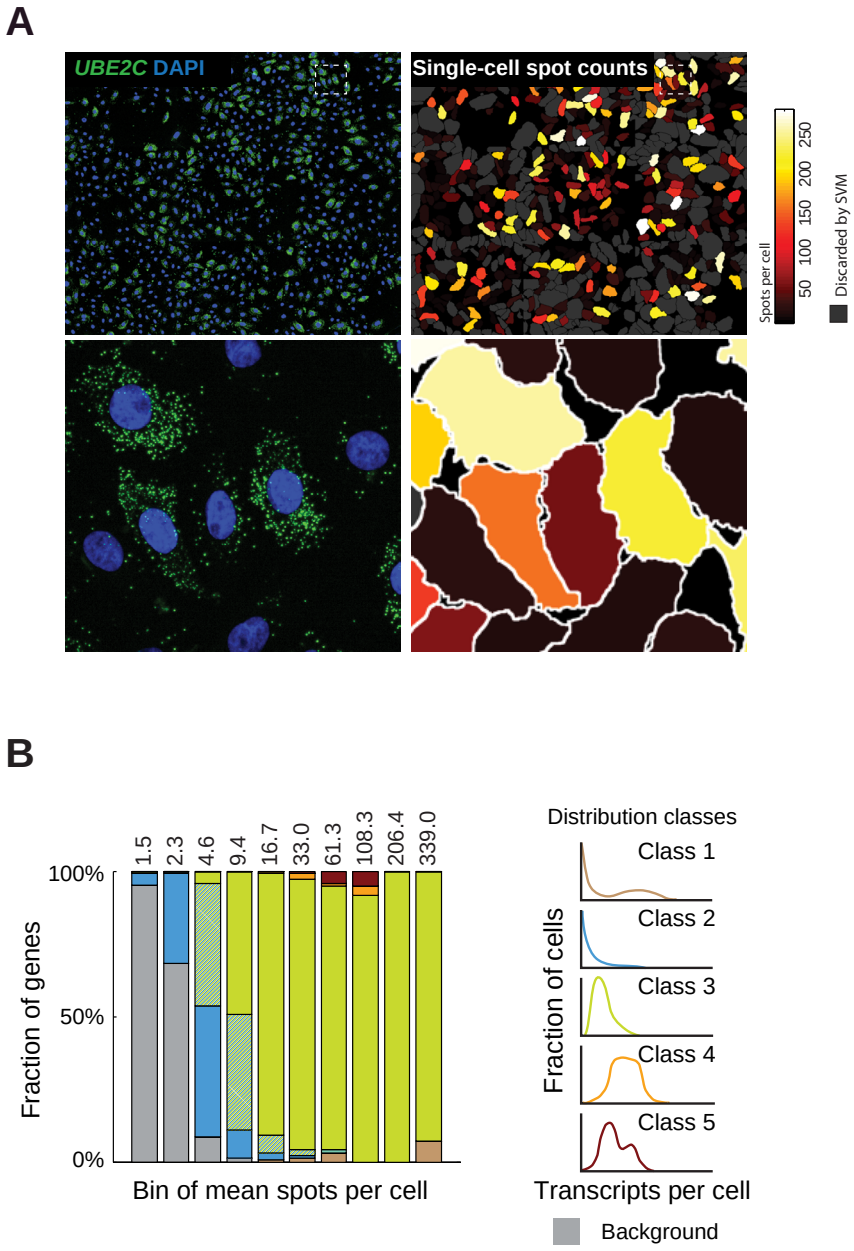
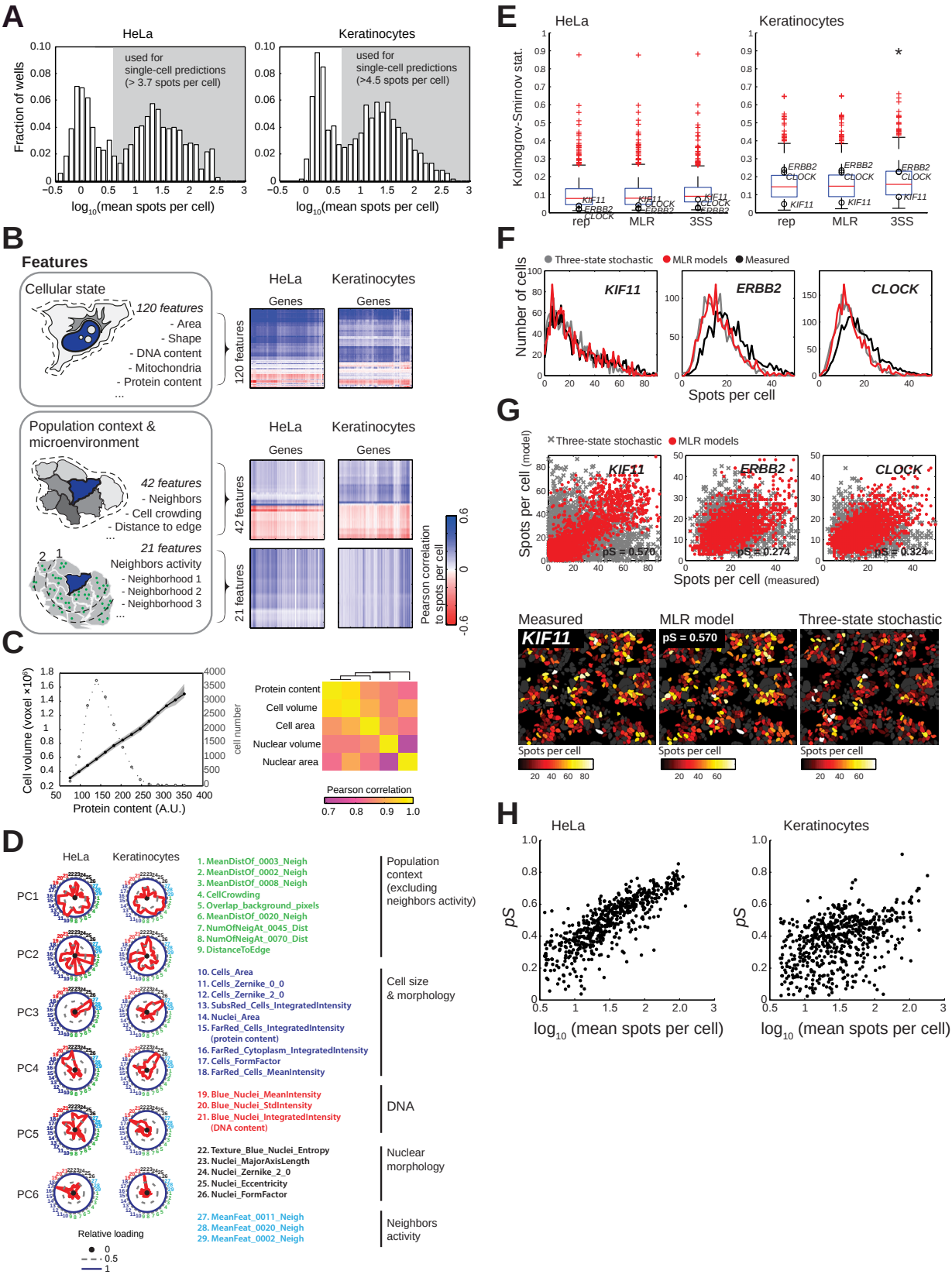


Figure S2



Click here to download Supplemental Figure FigS3_Limit.pdf

Figure S3

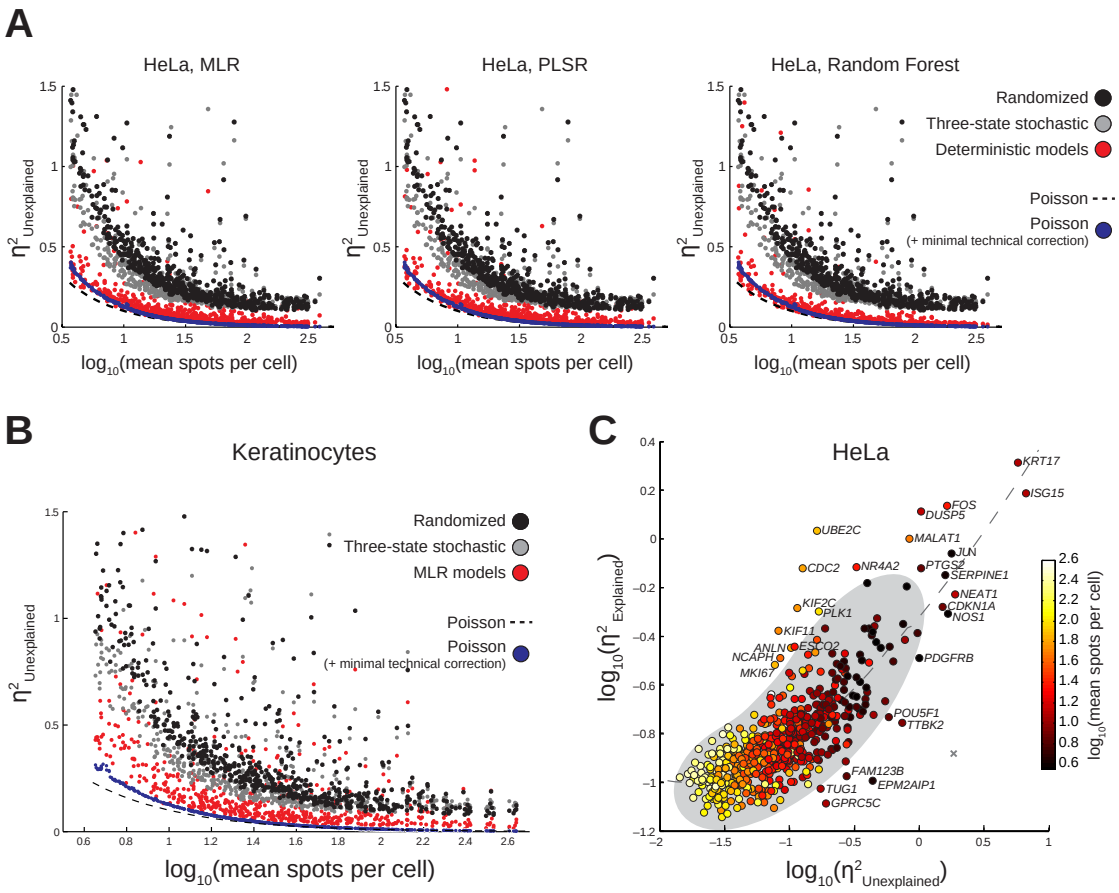


Figure S4

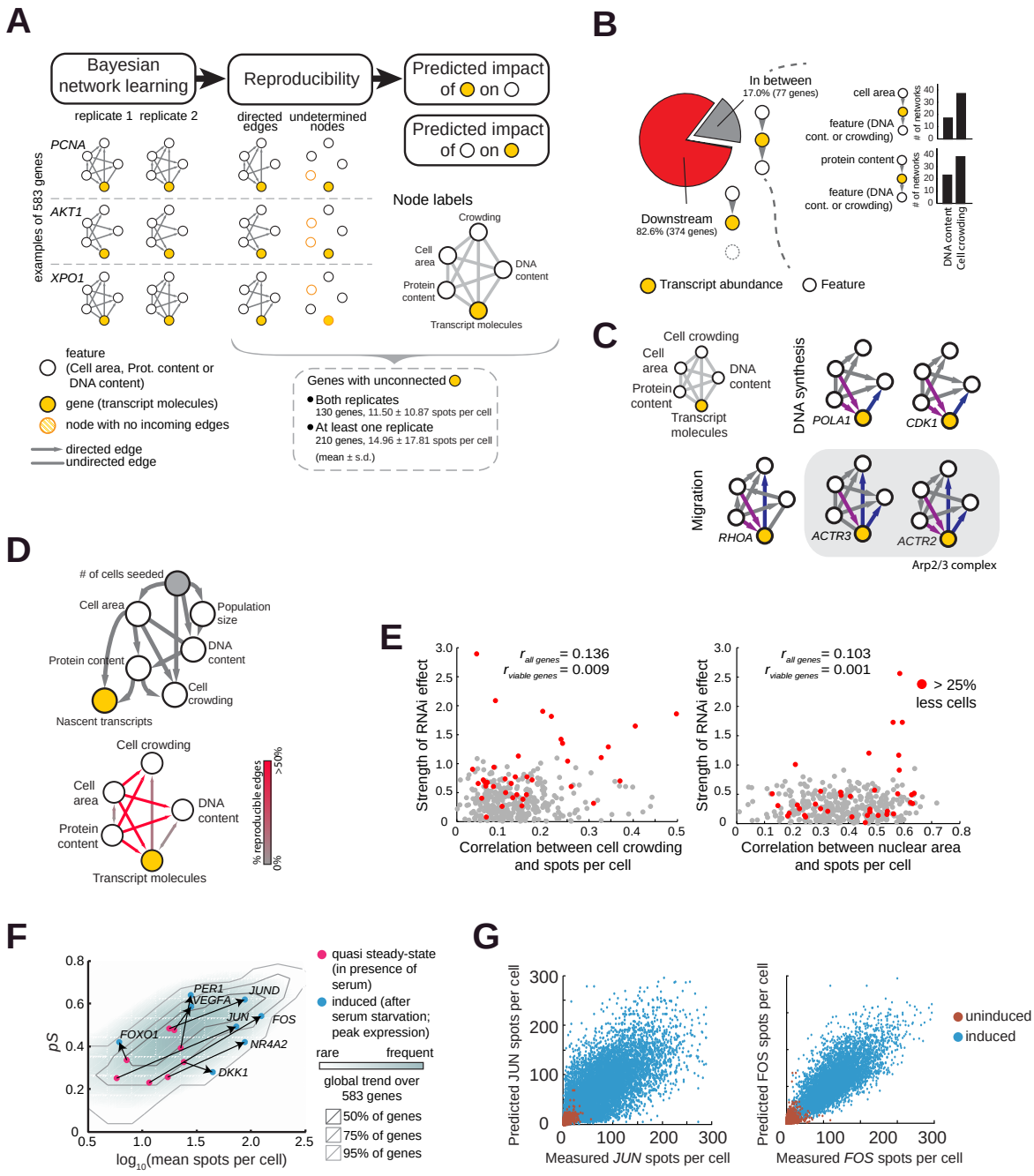

A
Bayesian network learning
Reproducibility
Predicted impact of on
Predicted impact of on
replicate 1 replicate 2
directed edges
undetermined nodes
examples of 583 genes
PCNA
AKT1
XPO1
Node labels
Crowding
Cell area
DNA content
Protein content
Transcript molecules
feature (Cell area, Prot. content or DNA content)
gene (transcript molecules)
node with no incoming edges
directed edge
undirected edge
Genes with unconnected
Both replicates
130 genes, 11.50 ± 10.87 spots per cell
At least one replicate
210 genes, 14.96 ± 17.81 spots per cell
(mean ± s.d.)
B
In between
17.0% (77 genes)
Downstream
82.6% (374 genes)
cell area
feature (DNA cont. or crowding)
protein content
feature (DNA cont. or crowding)
# of networks
DNA content
Cell crowding
Transcript abundance
Feature
C
Cell crowding
Cell area
DNA content
Protein content
Transcript molecules
DNA synthesis
POLA1
CDK1
Migration
RHOA
ACTR3
ACTR2
Arp2/3 complex
D
# of cells seeded
Cell area
Population size
DNA content
Protein content
Cell crowding
Nascent transcripts
Cell crowding
Cell area
DNA content
Protein content
Transcript molecules
% reproducible edges
>50%
0%
E
$r_{all\ genes}$ = 0.136
$r_{viable\ genes}$ = 0.009
Strength of RNAi effect
Correlation between cell crowding and spots per cell
$r_{all\ genes}$ = 0.103
$r_{viable\ genes}$ = 0.001
> 25% less cells
Correlation between nuclear area and spots per cell
F
pS
log10(mean spots per cell)
PER1
VEGFA
JUND
JUN
FOS
FOXO1
NR4A2
DKK1
quasi steady-state (in presence of serum)
induced (after serum starvation; peak expression)
rare
frequent
global trend over 583 genes
50% of genes
75% of genes
95% of genes
G
Predicted JUN spots per cell
Measured JUN spots per cell
Predicted FOS spots per cell
Measured FOS spots per cell
uninduced
induced

Figure S5

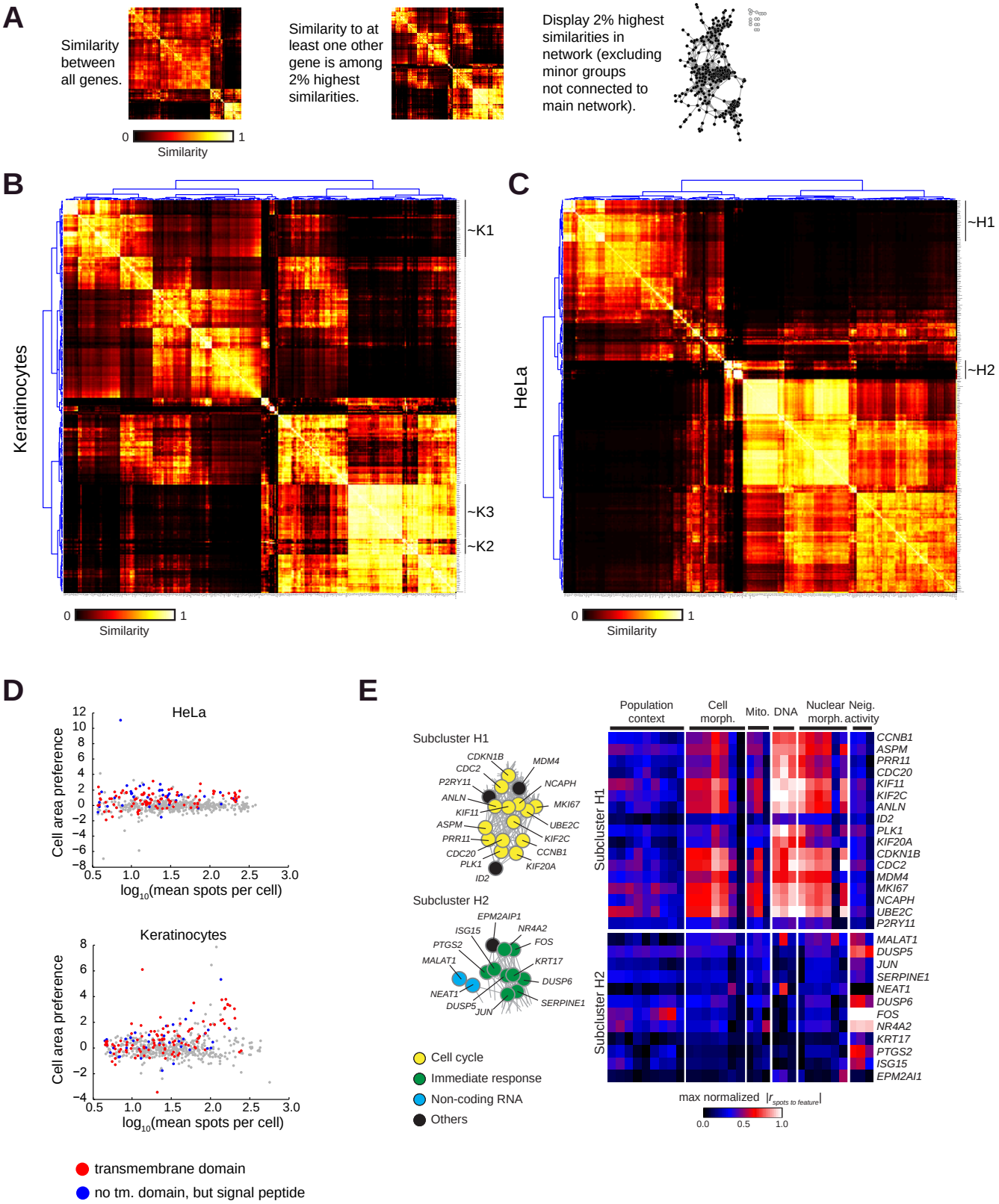
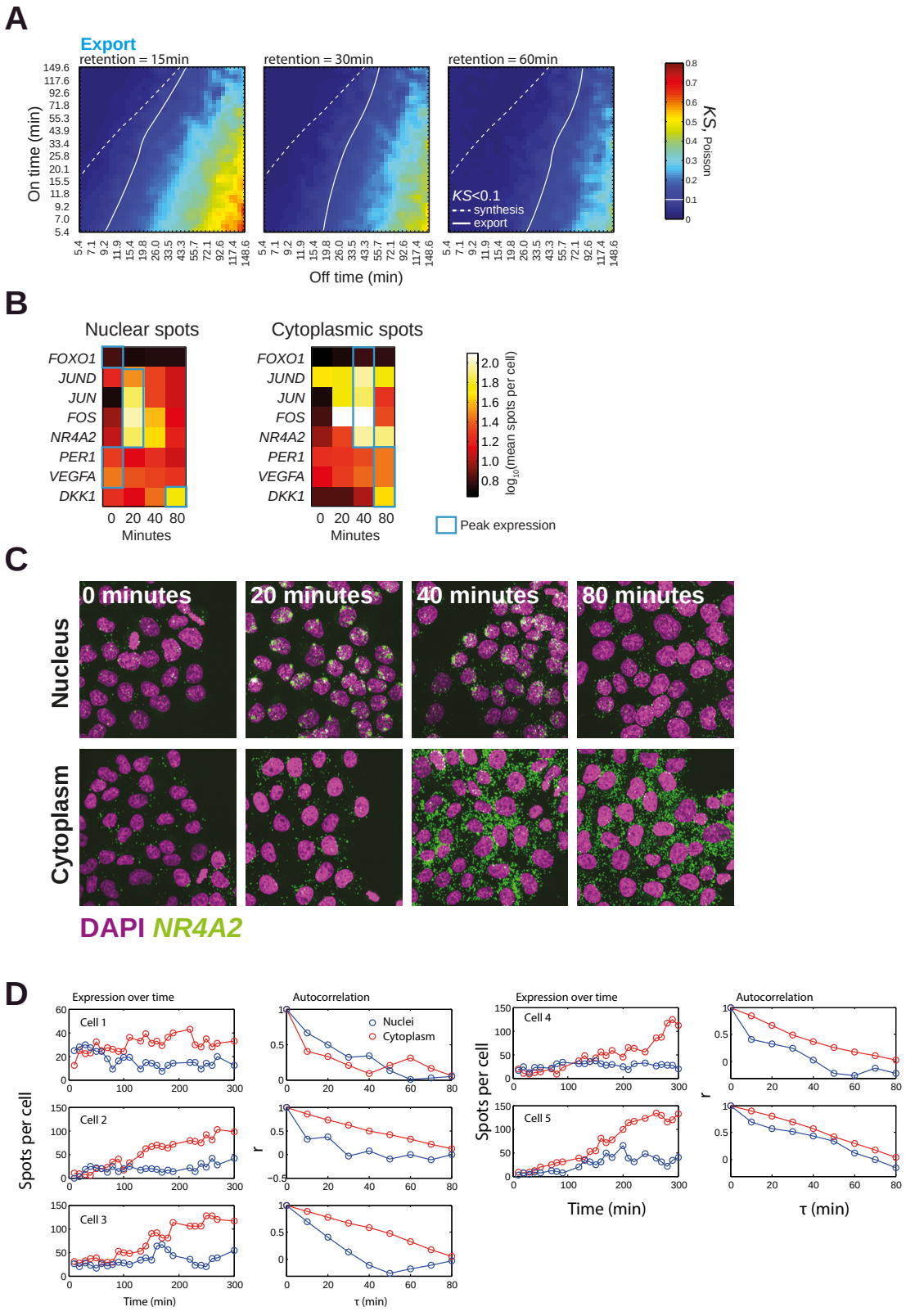


Figure S6



Acknowledgments

I would like to first acknowledge Lucas for inviting me to work in his lab. I thank Thomas Stoeger, who worked close with me in the project and produced about half the material presented in the chapters five to seven. I thank Mathieu for letting participate in his work, and Berend and Yauhen for help with computational infrastructure. I also thank Berend and Prisca for teaching me Matlab and image analysis when I started my PhD. I would like to mention Rene for help with experiments and the student James for assistance. J. Wilbertz (Friedrich Miescher Institute), J. Chao (Friedrich Miescher Institute), J. Ellenberg (European Molecular Biology Laboratory), E. Reichmann (University of Zurich) and L. Pontiggia (University of Zurich) provided reagents. Special thanks to Frank and Doris for countless evenings in the bouldering gym, and Doris again for helping me translate the Abstract into good German and commenting on detail the Introduction of the thesis. I finally thank all members of the lab for useful comments on the different manuscripts and passionate discussions on science.

i. Appendix. Curriculum vitae

BATTICH
Nicolas

DoB: 16th of June 1984
Nationality: Italian, Argentine

Education:

- Instituto Técnico Lorenzo Massa. S. M. de Tucumán Argentina, 2002.
  Técnico electromecánico. (Technical High School Diploma, Argentina).

- Stow College. Glasgow, Scotland. 2005.
  Access Course to Sciences (High School Diploma, Scottish A levels).

- The University of Glasgow. Glasgow, Scotland. 2010.
  Masters in Sciences, Molecular and Cell Biology – with Honors of the First Class.
  Master thesis title: *Biliary transporter function and the effects of cholestatic compounds in rat hepatocyte sandwich cultures.*

- PhD student at Prof. Dr. Lucas Pelkmans laboratory since August 2010 (ETH, Zurich).

- PhD student at the University of Zurich since January 2011.

Publications:

- Stoeger T*, **Battich N***, Herrmann MD, Yakimovich Y, Pelkmans L. Computer vision for image-based transcriptomics. *Methods*. 2015 May 23. pii: S1046-2023(15)00209-1. doi: 10.1016/j.ymeth.2015.05.016.

- Frechin M, Stoeger T, Daetwyler S, Gehin C, **Battich N**, Damm EM, Stergiou L, Riezman H, Pelkmans L. Cell-intrinsic adaptation of lipid composition to local crowding drives social behaviour. *Nature***.** 2015 Jul 2;523(7558):88-91. doi: 10.1038/nature14429.

- **Battich N***, Stoeger T*, Pelkmans L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nat Methods*. 2013 Nov;10(11):1127-33. doi: 10.1038/nmeth.2657.

- de Vos MG, Poelwijk FJ, **Battich N**, Ndika JD, Tans SJ. Environmental dependence of genetic constraint. *PLoS Genet*. 2013 Jun;9(6):e1003580. doi: 10.1371/journal.pgen.1003580.

  * Contributed equally

Poster presentations:

- 5th EMBO Meeting, (Amterdam,The Netherlands), 2013 .
- All SystemsX Meeting (Bern, Switzerland), 2012.
- IMLS, Institute Retreat (Switzerland), 2012.

Invited talks:

- EMBO Conference. From Functional Genomics to Systems Biology, 2014. Heidelberg, Germany.
- Statistical Methods for Post Genomic Data, 2015. Munich, Germany.
- EMBO Conference. Cellular Heterogeneity, 2015. Heidelberg, Germany.
- FranceBioImaging consortium's workshop on Bioimage-informatics, 2015. Paris, France.